8-2017

Uses of the Hypergeometric Distribution for Determining Survival or Complete Representation of Subpopulations in Sequential Sampling

Brooke Busbee

Stephen F Austin State University, brooke.busbee.616@gmail.com

Uses of the Hypergeometric Distribution for Determining Survival or Complete Representation of Subpopulations in Sequential Sampling

Creative Commons License

USES OF THE HYPERGEOMETRIC DISTRIBUTION FOR DETERMINING SURVIVAL OR COMPLETE REPRESENTATION OF SUBPOPULATIONS IN SEQUENTIAL SAMPLING

By

BROOKE BUSBEE, Bachelor of Science

Presented to the Faculty of the Graduate School of

Stephen F. Austin State University

In Partial Fulfillment

Of the Requirements

For the Degree of

Master of Science

STEPHEN F. AUSTIN STATE UNIVERSITY

August 2017

Uses of the Hypergeometric Distribution for Determining Survival or Complete Representation of Subpopulations in Sequential Sampling

By

Brooke Busbee, Bachelor of Science

APPROVED:

_____
Dr. Gregory K. Miller, Thesis Director

_____
Dr. Keith E. Hubbard, Committee Member

_____
Dr. Kent E. Riggs, Committee Member

_____
Dr. Chrissy J. Cross, Committee Member

_____
Richard Berry, D.M.A.
Dean of the Graduate School

Abstract

This thesis will explore the hypergeometric probability distribution by looking at many different aspects of the distribution. These include, and are not limited to: history and origin, derivation and elementary applications, properties, relationships to other probability models, kindred hypergeometric distributions and elements of statistical inference associated with the hypergeometric distribution. Once the above are established, an investigation into and furthering of work done by Walton (1986) and Charlambides (2005) will be done. Here, we apply the hypergeometric distribution to sequential sampling in order to determine a surviving subcategory as well as study the problem of and complete representation of the subcategories within the population.

Acknowledgements

When it comes to saying thank you, I always want my first thank you to go to The One who gave me life and breath and meaning, Jesus Christ. I pray that this document, the work I've done over the last two years, and the rest of my life are a testament to the faithfulness of God and serve to bring glory to Him alone.

Next, to my mom, my hero: thank you, thank you, thank you. You have shown me through all that you do, that hard work, dedication, and loving people are the key ingredients to success. Thank you for living a life full of love for your family, your community, and Jesus. No matter how many times I am asked the question, "Are you Cathy's daughter?" I will always be happy to answer yes.

To my dad: thanks for always pushing me to be the best I can be and do more than I thought I could. Thanks for encouraging me to major in math and pursue my masters, and for putting up with all my whining along the way.

To Mason: thanks for being everything I ever imagined a brother would be and more. You've been such an example of hard work, faith, endurance, and a true servant of the Lord to me over the past few years. I am so thankful to have spent this last year on the same campus.

To my grandparents, three of the greatest encouragers I've ever known. Though

only one of you have made this entire journey with me, this journey started years ago. It began with the hours spent reading books, the hours spent working on homework after school, and the investment placed in my education from even before I was born. To Mana especially, thank you for being a constant encouragement and example of humility, faith and service to the Lord. Thanks for believing I could finish this on days when I didn't even think I could myself.

To my graduate school family: I genuinely could not have done this without you. Thanks for keeping me sane, laughing with me, encouraging me, and both suffering and growing with me through these last two years together. God truly could not have blessed me with three better people to journey through grad school with. Lorna, Marissa, and Chad: I have learned so much from each of you. You have made these past two years not only bearable but two of the best, most profitable years of my life.

Occasionally God places people along our paths that really serve to point us in the direction He has intended for us. During my six years at SFA, I have been fortunate enough to have two of those people in my life: Dr. Beverly and Dr. Miller.

First, to Dr. Beverly: I don't think I could ever tell you thank you enough. Thank you for believing I could do this before I even believed in myself. Thank you for seeing something in me, and working to cultivate it from the first day I walked into your office. Thanks for not listening to me when I told you I never wanted to major in math… and then when I said I never wanted to get a Masters in math. Thanks for being there for me

every step of the way, from changing my major to math to defending my thesis. You have been such an encouragement to me and a wonderful example of humility and grace.

To Dr. Miller: Thank you for everything. I could not have made it through this process without you. Thank you for encouraging me to pursue my Masters and for working right alongside me to complete my thesis. It seems crazy to think that it's really done. Thanks for the many hours you have put into this project. You have served not only as my thesis advisor but also as an advisor and encourager in every area of my life over the last three years. You go above and beyond and truly exemplify working as if you are working for the Lord and not for men. You have taught me so much and helped me to grow as a student, as a thinker and as a communicator. You have answered my many questions and never passed up a teaching opportunity. I pray that as I move out of my role as a student and into my role as a teacher, I remember the passion you display in your teaching and the work you do to prepare and effectively communicate. It has been a true blessing to sit under your teaching and to work alongside you. I have learned more than I ever thought possible.

To the rest of my committee members, Dr. Hubbard, Dr. Riggs and Dr. Cross: you have each played a significant role in my education. Thank you first for serving on my committee and playing a role in this important phase of my life. Thank you also for investing in me, modeling a true love for education and your students, and teaching me both through your words and your actions.

Table of Contents

List of Figures

List of Tables

1. THE HYPERGEOMETRIC MODEL: SURVEY OF RESULTS

1.1: Introduction

This thesis explores different aspects of the hypergeometric distribution. The first part of the thesis looks at the origin of the distribution, its derivation, applications, properties of the distribution, relationships to other probability distributions, distributions kindred to the hypergeometric and statistical inference using the hypergeometric distribution. These topics have been summarized by evaluating various articles and other literature in order to synthesize information and organize it in a reasonable fashion.

The second portion of this thesis explores two particular sampling schemes that utilize variations of the hypergeometric distribution. These two schemes and the calculations involved are extensions of articles by G. S. Walton (1986) and C. A. Charlambides (2005). Both Walton and Charlambides look at sampling from populations that are divided up into distinct subcategories. Walton's work was evaluated then incorporated into a sequential sampling scheme that applies a sampling rule until only subcategory contains any occupants. While still examining subdivided populations, Charlambides' work was extended to take repeated samples until each subgroup contained one member only.

1.2: Origin of the Term 'Hypergeometric'

The word hypergeometric was first used to describe the hypergeometric function, rather than the hypergeometric distribution. In this section, the origin of the word "hypergeometric" will be explored along with the relationship between the original hypergeometric function and what is now known as the hypergeometric distribution. The hypergeometric mass function for the random variable X is as follows:

$$P(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}.$$

The hypergeometric distribution is used when the sampling of n items is conducted without replacement from a population of size N with D "defectives" and N-D "non-defectives" where x is the number of defectives found in the sample. Here the word defective can also be substituted with success depending on the scenario.

When searching through mathematical literature, it more common to come across articles dealing with the hypergeometric function as opposed to the hypergeometric probability distribution. While the hypergeometric function plays a role in the development of the hypergeometric distribution, the hypergeometric function was not the focus of this research.

The earliest uses of the word hypergeometric can be traced back to the middle 1600's. The name hypergeometric originated with Wallis in 1655 to the series whose n^{th} term is

$$a\{a + b\}\{a + 2b\} \dots \{a + (n - 1)b\}.$$

The word hypergeometric continued to be used in this sense until 1836 (Whittaker, Watson, 2010).

With the discovery of the hypergeometric function, the word hypergeometric became a widely used term in mathematical literature. The hypergeometric function can be used in many different contexts within mathematics, but it took years for the word hypergeometric to be used to describe a probability distribution. The origin of the term "hypergeometric distribution" is unclear in literature and though the name took off, it is not easily traced back to a single person, or time period.

In 1711, De Moivre solved "Huygens' fourth proposed problem" and his solution gave the probability of selecting x white and c-x black balls from a population of a white and b black balls, which gave rise to the hypergeometric mass function (Johnson, Kotz, 1992).

Twenty-nine years later, in 1740, Simpson derived the multivariate hypergeometric probability mass function. Still, after these two discoveries, it was not until 1843, when Cournot used the univariate hypergeometric mass function to describe certain situations within government, such as selection of parliament and juries, that the hypergeometric distribution gained a little more steam (Johnson, Kotz, 1992).

During the late 1800's and early 1900's, Karl Pearson began his work on limiting forms of discrete distributions. His study led him to investigate the properties of the hypergeometric distribution. Pearson's work was continued by Romanovsky on into 1925 (Johnson, Kotz, 1992).

Once the hypergeometric and multivariate hypergeometric distribution were well established, it did not take long for variations of the distribution to arise. These variations, such as the negative hypergeometric distribution, will be discussed in the future section titled "Kindred Distributions."

Since both the hypergeometric moment generating function and the probability generating function can be expressed in terms of the hypergeometric function, there is reason for the name "hypergeometric distribution." When referring to a random variable, probability distributions are often expressed in terms of their generating functions as opposed to their probability mass or density functions. In the early investigations, much of the work with the hypergeometric distribution was done, not by looking at the mass function, but instead looking at the moment and probability generating functions, written in terms of the hypergeometric function (Johnson & Kotz, 1992).

However, in this thesis, most of the exploration and connections are made by looking at the hypergeometric distribution or variations thereof in terms of the probability mass function, rather than in terms of the hypergeometric function.

1.3: Derivation and Elementary Applications

Derivation

The hypergeometric distribution is used when sampling without replacement from a finite, equally likely, population of items called defectives and non-defectives (or successes and failures). The hypergeometric distribution is often first explained by finding probabilities of drawing colored balls out of an urn or probabilities of particular

playing cards being drawn. In deriving the mass function of the hypergeometric distribution, consider a situation where a population includes both red and blue balls, all of the same size and weight (each equally likely to be chosen on any one draw). Suppose there are 100 total balls in a box, 30 of which are red and the other 70 are blue. Further, suppose it is desired that 10 balls be selected from the box and obtain 6 red and 4 blue balls. The total number of ways to select 10 balls from the original 100 is $\binom{100}{10}$. Similarly, it can be shown that there are $\binom{30}{6}$ ways to choose 6 red balls without replacement from the 30 total red, and $\binom{70}{4}$ ways to choose 4 blue balls from the 70 total blue in this population. Thus by the multiplication principle, the probability that in choosing 10 balls without replacement from a box of 100, 6 red and 4 blue are selected is given by:

$$P(6\ red\ selected\ out\ of\ 10) = \frac{\binom{30}{6}\binom{70}{4}}{\binom{100}{10}} = 0.03145.$$

Notice the above is just a fractional concept, where the numerator is the desired "part" or selection and the denominator is the "whole" or total number of possibilities.

To generalize the above, call the size of the population that is being selected from N, and the size of the sample being taken n. Similarly, call the number of red balls in the population D and the number of blue balls $N - D$. Now, say the desire is to select x, an arbitrary number, of red balls in our sample. Thus, if the sample size is given to be n, the

remaining $n - x$ balls will be blue. Next, by removing the words "red" and "blue" and viewing the red balls as defectives or successes and the blue balls as non-defectives or failures we have generalized the mass function of the hypergeometric distribution to

$$P(x\ defective\ in\ a\ sample\ of\ size\ n) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}.$$

Here, the domain of our mass function is such that x cannot be larger than D or n, whichever is smaller.

This, however, is not the only way to derive the mass function of the hypergeometric distribution. Instead of basing the derivation on a scenario, a more mathematical approach can be taken. This approach involves viewing the hypergeometric distribution through a "binomial lens." The hypergeometric distribution is characterized by the fact that the sample is taken without replacement. That being said, one can look at the hypergeometric sampling as a set of dependent Bernoulli trials where either successes or failures are chosen from the population of interest. Without loss of generality, say n items are sampled from a population of size N where the first x selections are "successes" (where there are D total successes in the population) and thus the final $n - x$ selections are "failures" (where there are N-D total failures in the population). Thus, the first x selections have probability:

$$\frac{D}{N} * \frac{D-1}{N-1} * \dots * \frac{D-(x-1)}{N-(x-1)} = \frac{D!}{(D-x)!} * \frac{(N-x)!}{N!}.$$

The last $n - x$ selections have probability:

$$\frac{N-D}{N-x} * \frac{N-D-1}{N-x-1} * \ldots * \frac{N-D-(n-x-1)}{N-x-(n-x-1)}$$

$$= \frac{(N-D)!}{\big(N-D-(\square-x)\big)!} * \frac{\big(N-x-(n-x)\big)!}{(N-x)!}.$$

Because the expressions above are for the very specific case of choosing all the successes first and all the failures last, all of the other ways to have x successes in n trials must also be considered, of which there are the combination $\binom{n}{x}$. Thus, synthesizing the above parts into a single expression, it can be seen that the probability of choosing x successes in n trials when choosing without replacement can be given by:

$$\binom{n}{x} * \frac{D!}{(D-x)!} * \frac{(N-x)!}{N!} * \frac{(N-D)!}{\big(N-D-(n-x)\big)!} * \frac{\big(N-x-(n-x)\big)!}{(N-x)!}$$

Rewriting the combination $\binom{n}{x}$ by applying its definition gives:

$$\frac{n!}{x!\,(n-x)!} * \frac{D!}{(D-x)!} * \frac{(N-x)!}{N!} * \frac{(N-D)!}{\big(N-D-(n-x)\big)!} * \frac{\big(N-x-(n-x)\big)!}{(N-x)!}$$

Then by combining terms and canceling out in order to see combinatorial expressions gives:

$$\frac{D!}{x!\,(D-x)!} * \frac{(N-D)!}{(n-x)!\,\big(N-D-(n-x)\big)!} * \frac{n!\,(N-n)!}{N!}$$

Last, by applying the definition of a combination once more, our expression for the hypergeometric mass function simplifies to:

$$\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}.$$

Notice that the above is equivalent to the probability mass function that was derived by considering the example earlier. For more on this approach, see Broca, 2008.

Duality and Symmetry

The hypergeometric probability distribution function can be alternatively represented by thinking of the probability of selecting x defectives out of a sample of size n rather than from the total number of defectives. There are $\binom{n}{x}$ ways for the x defectives to be distributed throughout the sample. This means, there are $\binom{N-n}{D-x}$ ways for the remaining defectives to be distributed through the rest of the population. Since the distribution of the defectives is now what is in question, the total number of ways for the D defectives to be distributed throughout the population of size N is $\binom{N}{D}$. Thus, the hypergeometric mass function can be alternatively expressed as:

$$\frac{\binom{n}{x}\binom{N-n}{D-x}}{\binom{N}{D}}.$$

This is referred to as the dual representation of the hypergeometric distribution probability mass function (Barnier, Jantosciak, 2002). By expanding the combinatorial terms and rearranging, it can be seen that the above expression is equivalent to the hypergeometric mass function. After rewriting the combinatorial terms and multiplying by the reciprocal of the denominator, we have:

$$\frac{n!\,(N-n)!}{x!\,(n-x)!\,(D-x)!\,(N-n-D+x)!} * \frac{D!\,(N-D)!}{N!}.$$

By reorganizing the above, we can get:

$$\frac{D!\,(N-D)!}{x!\,(D-x)!\,(n-x)!\,(N-D-n+x)!} * \frac{n!\,(N-n)!}{N!} = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}.$$

Thus, we see that the dual representation is in fact equivalent to the hypergeometric mass function.

Barnier and Jantosciak (2002) also show that the successes and the sample can be considered as two independent classifications and that gives rise to another representation of the probability mass function of the hypergeometric distribution. Thinking of the population as having two independent classifications, defective (or not) and sampled (or not), consider thinking of the intersection of the defective and sampled groups to give the probability of that intersection being of size "x." The number of total combinations of defectives and samples of size n are given by the following:

$$\binom{N}{D}\binom{N}{n}$$

Looking at the situation in terms of the categories of defective and sampled, we are interested in the intersection of defectives and sampled items. The entire population can be broken into four distinct groups by the classifications of interest.

Defective and Sampled of size x	Defective and not sampled of size D-x
Non-defective and Sampled of size n-x	Non-defective and not sampled of size N-n-D+x

Table 1: Group Sizes for Symmetric Representation

Using the multinomial coefficient, the probability mass function of the hypergeometric distribution can be written in the following way:

$$P(X = x) = \frac{\binom{N}{x, D-x, n-x, N-n-D+x}}{\binom{N}{D}\binom{N}{n}}.$$

The numerator is written compactly above, but the multinomial coefficient can be rewritten as

$$\frac{N!}{x!\,(D-x)!\,(n-x)!\,(N-n-D+x)!}.$$

This representation makes it easier to see that the symmetric representation is in fact equivalent to the hypergeometric mass function. To further show this relationship, we can multiply $\frac{N!}{x!(D-x)!(n-x)!(N-n-D+x)!}$ by the reciprocal of the combinations in the denominator to get:

$$\frac{N!}{x!\,(D-x)!\,(n-x)!\,(N-n-D+x)!} * \frac{D!\,(N-D)!\,n!\,(N-n)!}{N!\,N!}.$$

The above expression can be further simplified to

$$\frac{D!}{x!\,(D-x)!} * \frac{(N-D)!}{(n-x)!\,(N-D-n+x)!} * \frac{n!\,(N-n)!}{N!} = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}},$$

which is again the hypergeometric mass function.

Applications

The hypergeometric distribution is commonly studied in most introductory probability courses. In introducing students to the hypergeometric distribution, drawing balls from an urn or selecting playing cards from a deck of cards are often discussed.

Though not very advanced examples, they are simple and focus on sampling without replacement. They also provide a basis of understanding that students can build on and then further see how the hypergeometric distribution can be used to describe other situations.

Above, when discussing the derivation of the hypergeometric distribution, a univariate urn example was used where 10 balls were chosen from a mixture of 100 red and blue balls. This is probably the most widely used elementary metaphor for applications of the hypergeometric distribution.

The hypergeometric distribution can be used to calculate probabilities in a wide variety of scenarios. Fury, Batiwlwalla, Gregersen and Li (2006) used the hypergeometric distribution to investigate scenarios within gene selection. Schuster (1991) uses the hypergeometric distribution in his article entitled *The Statistician in a Reverse Cocaine Sting.* In this article Schuster describes a drug bust, the plan of action devised by the police and uses the hypergeometric distribution to find the probability that cocaine purchased in a reverse sting was in fact cocaine. Cooper (2016) even uses the hypergeometric distribution to establish strategies for winning a game of poker.

Many of the articles used in the development of this paper, are not exploring the hypergeometric distribution itself, but are studying topics in the sciences and see that the hypergeometric distribution is a result of the sampling scheme employed in the experiment or study. This leads to more complex applications of the hypergeometric

distribution and even to the development of other distributions related to the hypergeometric distribution, such as the negative hypergeometric distribution seen later.

1.4: Relationships to Other Probability Models

In this section, both approximations and families of distributions that the hypergeometric belongs to will be discussed. Well-known approximations to the hypergeometric distribution are given by the binomial distribution, the normal distribution and less commonly, the Poisson distribution. The hypergeometric distribution belongs also to several families of discrete probability distributions. These are discussed later in this section.

Approximations

It is often appropriate and useful to approximate the hypergeometric distribution. Most often, this is done by a Binomial approximation. The binomial distribution is very widely known and in some cases, it can result in a simpler calculation than that of the hypergeometric distribution. The binomial is used in situations when sampling occurs with replacement and the probability of x successes in n Bernoulli (success/fail) trials is desired. This differs from the hypergeometric in that the sample is taken with replacement versus without replacement in the hypergeometric case, and thus the chance of obtaining a "success" is constant on each draw. Using the traditional form of the hypergeometric mass function, define H as

$$H(x: N, n, D) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}.$$

Additionally, let the binomial mass function B be denoted

$$B(x: n. p) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Further note that $H(x; N, n, D) = H(x; N, D, n) = H(n - x; N, N - D, n) = H(D - x; N, N - n, D)$. Thus, there are four resulting forms of the binomial approximation:

$$B\left(x; n, \frac{D}{\square}\right) = B\left(x; D, \frac{n}{N}\right) = B\left(n - x; N - D, \frac{n}{N}\right) = B\left(D - x; N - n, \frac{D}{N}\right)$$

(Brunk, Holstein, Williams, 1968). As N and D approach infinity and $\frac{D}{N}$ approaches p, the limit of the hypergeometric mass is given by

$$\lim_{N,D \to \infty, \frac{D}{N} \to p} \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} = \binom{n}{x} p^x (1-p)^{n-x}$$

(Miller, 2006).

Plachky (2003) considers the random variables X_1 and X_2 that are each, independently binomially distributed with sample size n and probability p. He then shows that the conditional distribution of X_1 given the sum of X_1 and X_2 ($X_1 + X_2 = y$) is hypergeometric with parameters $n_1 + n_2$, n_1, and y. He notes also that as the sum of the sample sizes approaches infinity the hypergeometric distribution converges to the binomial distribution. Similarly, Plachy also shows that the same relationship exists between the negative binomial distribution and the negative hypergeometric distribution.

Lastly, Sandiford (1960) presents a "new" binomial approximation to the hypergeometric distribution. Sandiford develops this new binomial approximation by

equating the mean and variance of the binomial distribution to that of the hypergeometric distribution. This new approximation is skewed roughly as much as the hypergeometric distribution and is skewed in the same direction, providing what Sandiford believed to be a better fit when compared to other approximations.

The normal distribution is also used to approximate the hypergeometric distribution. The normal probability distribution is commonly used as an approximating distribution for large sample sizes due to the invocation of the central limit theorem. As the sample size of the hypergeometric distribution increases, along with the number of defectives and the population size increasing, the normal distribution becomes an appropriate approximating distribution. In fact, as the sample size grows, the normal approximation becomes more and more accurate and the margin of error decreases.

An issue of concern among those approximating the hypergeometric is: which approximation is best? Ling and Pratt (1984) concluded that among 15 approximations, the Peizer approximation is best. The Peizer approximation involves transforming a two-by-two table and calculating a normal deviate. On the other hand, Liberman and Owen (1961) consider the binomial approximation with the smallest sample size to be best. Nicholson (1956) uses results from Feller's normal approximation to the binomial to find a normal approximation to the hypergeometric distribution. He notes that as the sample size grows the absolute error decreases.

Along with approximating the standard hypergeometric distribution, variations of the hypergeometric distribution can also be approximated. Terrapabloran (2011) gives an

approximation to the negative hypergeometric distribution, which will be discussed in the section discussing kindred distributions, using the Poisson distribution. Childs and Balakrishanan (2000) give a method for approximating the multivariate hypergeometric distribution. Childs and Balakrishanan use continuous random variables to develop their approximations. They show that their approximation methods could be used in testing hypotheses concerning the parameters of the multivariate hypergeometric distribution.

Families

Probability distributions can be classified in broad categories, known as families of distributions, based on characteristics and properties of the distribution. The hypergeometric belongs to several families of discrete distributions.

Ollero and Ramos (1995) show the hypergeometric distribution to be a member of a subfamily the Pearson family of distributions. The Pearson family is a very broad classification of distributions developed by Karl Pearson. Ollero and Ramos conjecture that the Pearson family consists of a subgroup of distributions, including the hypergeometric that are "generalized-binomial" distributions. They define "generalized-binomial" to mean that the distribution is that of the number of successes among independent trials where the success probability is not the same for every trial.

The hypergeometric distribution also belongs to a generalized family of distributions described by Mathai and Saxena (1967). This generalized family is defined by using the hypergeometric function and Mathai and Saxena note that this family encompasses almost all classical probability distributions.

Another family of distributions is one generated by the bivariate Bernoulli distribution. Marshall and Olkin (1985) show that the multivariate hypergeometric distribution is also a part of this family. This distribution will be further discussed in a later chapter.

1.5: Properties of the Hypergeometric Distribution

In this section, various properties of the hypergeometric distribution will be explored. First we examine properties relating to the moments of the hypergeometric distribution and then move to particular applications of these properties.

Moments

When considering the moments of a probability distribution, it can be very helpful to know the moment generating function (MGF) of the distribution. The MGF provides a concise way to represent all of the moments of a probability distribution.

Johnson and Kotz (1992) give several different approaches to finding the moments of the hypergeometric distribution. The first they look at is the factorial moment generating function, which can be used to find the moments of the distribution. They also show that the moment generating function can be represented using differential equations and the hypergeometric function. This is one more reason the hypergeometric distribution bears the name hypergeometric. It is also noted that the moment generating function of the multivariate hypergeometric distribution can be represented in a similar fashion. However, Lessing (1973) shows that the MGF of the hypergeometric can be expressed without using the hypergeometric function.

When studying probability distributions, it is often of utmost importance to understand and be able to calculate basic summary properties of the distribution such as the mean and the variance. More specifically, if the moments of the distribution can be found, so can the mean and variance. Recall, the probability mass function of the hypergeometric distribution is given by:

$$P(x) = P(X = x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}.$$

The domain of the hypergeometric mass function consists of integers from 0 to the minimum of D and n. This is because even if D<n, there can be at most D defectives (successes) in a sample of size n. Kemp and Kemp (1956) highlight the fact that the hypergeometric mass function can also be expressed as a product of a constant and the hypergeometric series. The focus here, will be on the mass function expressed in terms of combinations rather than the hypergeometric function. To find the expected value (mean) of the hypergeometric distribution, use the definition of expected value to produce the following:

$$E(X) = \sum xP(x)$$

$$E(X) = \sum x\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

From this point, it is advantageous to expand the combinatorial terms and attempt to produce a second hypergeometric mass function inside the original, so that once constants have been removed, the sum will equal to 1.

$$E(X) = \sum \frac{[D!\,(N-D)!]/[x!\,(D-x)!\,(n-x)!\,(N-D-n+x)!]}{N!/[n!\,(N-n)!]}$$

Now, by the definition of factorial, we can rewrite D!, x!, N!, and n! by pulling out the first term of the factorial and multiplying by the factorial represented by the remaining terms. This will create of the sum of a hypergeometric probability mass function after pulling the constants out of the original sum. The remaining steps are shown below:

$$E(X) = \sum \frac{[D(D-1)!\,(N-D)!\,x]/[x(x-1)!\,(\square - x)!\,(n-x)!\,(N-D-n+x)!]}{[N(N-1)!]/[n(n-1)!\,(N-n)!]}$$

$$E(X) = n\frac{D}{N}\sum \frac{[(D-1)!\,(N-D)!]/[(x-1)!\,(D-x)!\,(n-x)!\,(N-D-n+x)!]}{[(N-1)!]/[(n-1)!\,(N-n)!}$$

$$E(X) = n\frac{D}{N}\sum \frac{\binom{D-1}{x-1}\binom{N-D}{n-x}}{\binom{N-1}{n-1}}$$

Because the last sum above fits the form of the hypergeometric pmf, it equates to 1 also and we have that:

$$E(X) = n\frac{D}{N}.$$

The expected value of the hypergeometric can also be derived by creating an indicator function, I_k, such that I_k is 1 if the kth item sampled is a defective and 0 if the kth item sampled non-defective. Thus, because the hypergeometric distribution models the number of defectives in a sample of size n, the expected value can be represented as:

$$E(X) = E\left(\sum_{k=1}^{n} I_k\right)$$

Because expected value is a linear operator, the expected value operator on the right hand side can be moved inside of the sum to create the following expression:

$$E(X) = \sum_{k=1}^{n} E(I_k)\,.$$

$$E(X) = \sum_{k=1}^{n} \frac{D}{N}$$

This is because the expected number of defectives in a sample is equal to the proportion of defectives in the entire population of N items. That is, unconditionally the chance that the k^{th} sampled item is "defective" is $\frac{D}{N}$ (Miller, 2006). Finally, because the above is a sum of a constant, we can represent the expected value of the hypergeometric as $E(X) = n\frac{D}{N}$, which is the same conclusion reached by using the definition and combinatorial identities above.

Looking now at the variance of the hypergeometric distribution, recall that the variance of a random variable X is given as $Var(X) = E(X^2) - [E(X)]^2$. Similar to the methods used above to derive the expected value, the variance can also be derived in two ways.

First, using the first factorial moment to calculate the variance, consider:

$$E[X(X-1)] = \sum x(x-1)\frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}$$

$$= \sum x(x-1)$$

$$* \frac{[D(D-1)(D-2)!\,(N-D)!]/[x(x-1)(x-2)!\,(D-x)!\,(n-x)!\,(N-D-n+x)!]}{[N(N-1)(N-2)!]/[n(n-1)(n-2)!\,(N-n)!]}$$

$$= n(n-1)\frac{D}{N}\frac{D-1}{N-1}\sum \frac{\binom{D-2}{x-2}\binom{N-D}{n-x}}{\binom{N-2}{n-2}}$$

$$= n(n-1)\frac{D}{N}\frac{D-1}{N-1}$$

From this calculation and that of the expected value earlier, the variance can be found. The above gives us an expression for $E(X^2 - X) = E(X^2) - \mathrm{E(X)}$. We can then isolate $E(X^2)$ to produce

$$E(X^2) = n(n-1)\frac{D}{N}\frac{D-1}{N-1} + \mathrm{n}\frac{D}{N}.$$

Using the fact that $Var(X) = E(X^2) - [E(X)]^2$, we get

$$Var(X) = n(n-1)\frac{D}{N}\frac{D-1}{N-1} + n\frac{D}{N} - \left(n\frac{D}{N}\right)^2$$

$$= n(n-1)\frac{D(D-1)}{N(N-1)} + n\frac{D}{N}(1-n)$$

$$= n\frac{D}{N}\frac{N-D}{N}\frac{N-n}{N-1}.$$

The indicator function approach used to find the expected value could also be used to calculate the variance of the hypergeometric distribution. Using once more the fact that $Var(X) = E(X^2) - [E(X)]^2$, we can see:

$$Var(X) = E\left[\left(\sum_{k=1}^{n} I_k\right)^2\right] - \left(\frac{D}{N}\right)^2.$$

Calculations similar to those done previously when finding the expected value will produce the same expression for Var(X). Similar methods could be used to find other moments of the hypergeometric distribution, but for most practical purposes, the mean and variance are sufficient summary statistics.

Remembering the duality and symmetry of the hypergeometric distribution discussed earlier, these properties of the distribution allow for alternative representations of the hypergeometric mass function and can also be counted as significant properties of this distribution. Bol'shev (1964) highlights this in his paper discussing simple sequential sampling schemes. He states that the hypergeometric distribution mass function is "invariant to permutation of the rows or columns of the 2 by 2 array," which results in the fact that the hypergeometric mass function can be written in multiple equivalent ways and the moments can be found by using any of these representations of the mass function.

Various Other Properties

Godbole (1990) shows that the distribution of the number of success runs, in a fixed number of trials, of the hypergeometric distribution is hypergeometric itself. He then goes on to find the distribution of the longest success runs and the distribution of the waiting time until the r^{th} success run of a particular length. The distribution of the waiting time until the r^{th} success run of a particular length is related to the negative hypergeometric distribution. Guenther (1978) further discusses the applicability of the

hypergeometric distribution when dealing with success runs. Guenther's "runs test" is discussed in later in the section titled "Statistical Inference."

Joarder (2012) looks at particular probabilities of interest when studying the hypergeometric distribution. He examines things such as the probability that a particular element of the population is included in a particular selection and situations where the sample space is not equally likely. He provides identities and theorems to validate his work. Joarder gives various examples of calculations and demonstrates how basic statistical properties and properties specific to the hypergeometric can be used to solve problems.

Algorithms Associated with Calculations

When calculating probabilities, especially in more complex cases, it is often beneficial to ask if there is a way to make the computations easier. In the present day, with very little knowledge of programming or mathematical software, the average mathematician or statistician could pool resources and use modern software to perform complex computations rather quickly with ease. This is due to the fact that, in the past, complex algorithms were generated so that computations could be performed more simply using the appropriate software. Calculations of this nature can be seen in articles such as that of Alvo and Cabilio (2000). Other properties not included in these subsections are summarized in Johnson and Kotz (1992).

1.6: Kindred Hypergeometric Distributions

In this section, some variations of the hypergeometric distribution will be discussed. The *negative* (sometimes referred to in literature as the inverse) hypergeometric, the compound hypergeometric, the noncentral hypergeometric, the multivariate hypergeometric, as well as other notable distributions will be discussed in this section. The variations discussed here cover the popular variants of the hypergeometric distribution, but the list of variants discussed here is not exhaustive.

The Negative Hypergeometric Distribution

Just as the binomial and the negative binomial distribution are related, so are the hypergeometric and the negative hypergeometric distributions. The hypergeometric probability mass function gives the probability of x successes being chosen in n trials, when the sampling is done without replacement. The negative hypergeometric gives the probability that it would take a random number of trials x to observe a fixed number of successes say r, still sampling without replacement. When using the hypergeometric distribution, the probability a random number of successes occur in a fixed number of draws is the desired quantity. Thus, it can be seen that the negative hypergeometric distribution fixes the number of successes and allows the number of trials to be the random variable. The probability mass function for the negative hypergeometric distribution is

$$P(X = x) = \frac{\binom{x-1}{r-1}\binom{M-x}{N-r}}{\binom{M}{N}}.$$

Here, x denotes the number of items that need to be sampled in order to observe r successes/defectives. The total number of possible successes/defectives in the population is N and M is the size of the entire population being sampled from. The negative hypergeometric distribution can be extended to the multivariate situation as well. The multivariate negative hypergeometric will be used later when exploring the extensions of Walton (1986).

Just as in the case of the negative binomial, the first r-1 successes will be observed in x-1 trials and the r^{th} success is observed on the x^{th} trial. Because the items in the sample are being taken without replacement, the probability of the first r-1 successes occurring in x-1 trials has hypergeometric probability

$$\frac{\binom{N}{r-1}\binom{M-N}{x-r}}{\binom{M}{x-1}}.$$

Then, the probability that the r^{th} success occurs on the x^{th} trial is given by:

$$\frac{N-(r-1)}{M-(x-1)}.$$

Combining the two generates the probability mass function above (Miller and Fridell, 2007).

Once again, just as the hypergeometric distribution can be approximated by the binomial distribution, it is noted in Schuster and Sype (1987) that for large values of N, the negative hypergeometric can be approximated by the negative binomial distribution. Schuster and Sype also calculate the expected value of the negative hypergeometric

distribution to be $E(X) = r/\binom{n+1}{N+1}$ and through a series of relationships between the waiting time until the first success also derives a formula for the variance of the negative hypergeometric distribution. Lastly, Schuster goes on to demonstrate how to estimate the parameters of the negative hypergeometric distribution.

D'Elia (2003) discusses a unique application of the negative hypergeometric distribution to ranking. She works to analyze rank data to both study preferences of consumers and the satisfaction of consumers. The hypergeometric distribution is used to perform both of these analyses. Instead of looking at the ranking of preferences as a multivariate problem, D'Elia considers the ranks of the items in reference to one particular item to maintain a univariate scenario. This can be done by employing a multistage or iterated sampling structure where the first item chosen is the most preferred and this process is repeated until only the least preferred item remains. D'Elia goes on to show how to find the moments of this specific negative hypergeometric distribution, methods for parameter estimation and covers an in depth example of this distribution being used to study preferences in the olive oil market.

Compound Hypergeometric Distribution

The compound hypergeometric distribution was defined by Hald (1960). When working with compound distributions in order to find the probability mass function of a random variable X, the joint probability mass function of X and Y must first be found. This is because compound distributions involve conditioning, and conditional mass functions are a function of the joint mass function. The first step in developing the

compound hypergeometric is to find the joint mass function of X and Y. Hald defines the joint mass function to be

$$P(X,Y) = f_s(w+y)\frac{\binom{n}{w}\binom{N-y}{y}}{\binom{N}{w+y}},$$

where N is the population size, n is the sample size, X is the number of defectives in the population, W is the number of defectives sampled, and $Y = X - W$ is the number of defective items not sampled. Thus, in this case $X = W + Y$. Note also that, f_s gives the probability that a population of N items contains X defectives and the conditional distribution $P(W|X)$ is the probability that a sample of n items contains w defectives given that the population contains a total number of X defectives. The conditional hypergeometric distribution $P(W|X)$ is as follows:

$$P(W|X) = \frac{\binom{X}{w}\binom{N-X}{n-w}}{\binom{N}{n}}.$$

The joint distribution of X and W can also be found to be:

$$P(X,W) = f_s(X)P(W|X)$$

and from the above the marginal distribution of W can be found to be:

$$g(w) = \sum_{w} P(W|X)f_s(X)$$

and gives the average probability of x defectives being sampled.

Using the information above, the marginal distribution of W, can be given by:

$$g(w) = \binom{n}{w} \sum_{y=0}^{N-n} f_s(w+y) \frac{\binom{N-n}{y}}{\binom{N}{w+y}}.$$

A random variable having the above mass function is defined to be a compound hypergeometric variable. It is named such because it is generated by averaging the distribution of W over all possible values of $X = w + y$.

This creates a situation where there are two random variables. Both the number of successes that will be drawn (W) and the number of successes that will be left in the population (Y) are unknown. Hald further develops this distribution by calculating the moments of (W,Y) by first computing factorial moments, conditional expectation and variance.

Noncentral Hypergeometric Distribution

When referring to a distribution as a noncentral hypergeometric distribution it is beneficial to ask if the distribution is of the Wallenius noncentral hypergeometric distribution type or what is known as Fisher's noncentral hypergeometric distribution. Fisher's noncentral hypergeometric distribution is sometimes referred to as the extended hypergeometric distribution.

The multivariate Wallenius noncentral hypergeometric distribution is used in cases where each selection taken without replacement and is not equally likely, or in other words, when the sampling is biased. For instance, consider that black and white balls were being drawn from an urn, but the white balls were two times larger than the black balls. This situation could be modeled using the Wallenius noncentral

hypergeometric distribution. More generally speaking, the population consists of N balls of c different colors and n balls are to be sampled from the population. In this situation, the probability that a particular ball is sampled is proportional to the "weight" of the ball, where the weight depends only on the color of the ball (Fog, 2008). Though not stated here, the mass function also appears in Fog (2008).

The multivariate Fisher's noncentral hypergeometric distribution, sometimes called the extended hypergeometric distribution, can be used in simulation type experiments where the outcome is known beforehand. For instance, knowing row or column totals in a contingency table prior to filling the cells with observations. Accordingly, Fisher's noncentral hypergeometric distribution is used for various tests when analyzing contingency tables. Liao (1992) discusses different applications of the extended hypergeometric distribution for 2 by 2 contingency tables and derives the mean and variance of Fisher's noncentral hypergeometric distribution. The multivariate Fisher's noncentral hypergeometric distribution is defined by Harkness (1965) as the "conditional distribution of independent binomial variates given their sum." Harkness also notes that there exists a binomial approximation for Fisher's noncentral hypergeometric distribution. Elsinga and Peizer (2011) give saddlepoint approximations for the extended hypergeometric distribution.

In both the Wallenius and Fisher cases, if n=1 the distribution reduces to the multinomial distribution. Also, if all weights are equal, the distributions reduce to the multivariate hypergeometric distribution.

Other Notable Distributions

Irwin (1954) discovered in his study of the spread of infectious disease that the spread of an infectious disease through a household could be modeled by a variant of the hypergeometric distribution. Irwin (1975) develops a distribution titled the Generalized Waring Distribution, which in one particular case reduces to what is termed the Yule Distribution, another variant of the standard hypergeometric model.

1.7: Statistical Inference

In this section estimation of the hypergeometric parameters and statistical tests that have hypergeometric sampling distributions will be discussed.

Parameter Estimation

In many statistical problems, it is often advantageous to estimate the unknown parameters. There are several different methods used for estimation, with the most common being maximum likelihood estimation. Zhang (2009) uses the maximum likelihood method to estimate the population size, N, of the hypergeometric distribution. He finds that the maximum likelihood estimator for the population size of a hypergeometric distribution is given by the integer part of $\frac{Dn}{x}$, that is if the estimator is fractional, the estimator is rounded down to the nearest whole number. Zhang (2009) also finds a maximum likelihood estimator for the number of defectives in the population. His result shows that the maximum likelihood estimator for D, the number of defectives, is not unique but is given by either $\frac{x(N+1)-n}{n}$ or $\frac{x(N+1)-n}{n}+1$. It is important to note that, though not discussed, the method of moments approach to estimation can be used for the

hypergeometric distribution as well. Method of moments estimation involves simply equating the population moments to the expected values of the moments.

Other methods of estimation are explored in Tohma (1991). In his paper, Tohma examines six different methods for estimating the parameters of the hypergeometric distribution and discusses their accuracy. The six methods explored are: differencing, Stirlings formula, the binomial distribution, a combination of the binomial distribution and Stirlings formula, the normal distribution and the least squares sum method. He concludes that the best estimation method is the least squares sum method when estimating the population size.

Along with estimating parameters, it is sometimes nice to be able to develop confidence intervals for those estimates. Thompson (2012) discusses general methods for both estimating parameters and developing confidence intervals for those estimates. Wang (2015) elaborates on the ideas established by Thompson and notes that the unknown parameter in the hypergeometric distribution could either be the population size, N, or the number of defectives in the population, which Wang denoted as M. (Defined as D in the development of this thesis.) Wang notes that Thompson develops approximate intervals that can be unreliable. Wang argues for the use of exact intervals and goes on to show the development of an algorithm to allow for exact confidence intervals for the population size and number of defectives.

Tests with Hypergeometric Sampling Distributions

When performing any type of statistical test, it is important to identify both the test statistic and the associated sampling distribution of that test statistic. The hypergeometric distribution is used as the sampling distribution for various statistical tests, such as Fisher's exact test, a median test, runs test and other analyses of contingency tables.

Looking first at Fisher's Exact Test, which can be used to analyze two by two contingency tables comparing success probabilities, where the marginal totals (column and row totals) are known prior to collecting the data. Fisher's Exact Test can also be used when the marginal totals are unknown and in this case is viewed as a competitor to Pearson's χ^2 test for comparing two success probabilities. Fisher's Exact Test relates to the hypergeometric distribution because the sampling distribution for the test is hypergeometric where the parameters are the row and column totals, which are generally known before sampling (Hollander, Wolfe, Chicken, 2014).

Mehta and Patel (1983) give an algorithm for performing Fisher's Exact Test on r by c contingency tables. This is an extension of Fisher's Exact Test for tables of any dimension. Mehta notes that comparing Fisher's Exact Test and the χ^2 test can often lead to contradictory results.

Gart (1963) developed a median test based on an experiment where his application involves two groups of mice that are exposed to radiation where one of the groups had previously been exposed. One mouse in the control group was assigned to be

the "median" mouse and the group was observed until the median died. This led to the development of his median test. In this procedure, the median of each population must be found and the number of observations in the treatment group less than the "median" of the control group is the variable of interest, say X. Gart shows that the null or sampling distribution of X is given by:

$$P(x) = \frac{2s+1}{n+2s+1}\frac{\binom{n}{x}\binom{2s}{s}}{\binom{n+2s}{s+x}}$$

where s+1 corresponds to the placement of the median in the population. The sampling distribution developed by Gart is not entirely hypergeometric in nature, but it can be seen from the combinatorial structure that the distribution is related to that of the hypergeometric. Though there is a fractional coefficient in the sampling distribution that is not present in the standard hypergeometric mass function, the combinatorial terms in the numerator can be combined to form the combination in the denominator. This is the characteristic of the above distribution that relates it to the hypergeometric distribution.

Guenther (1978) discusses the use of the hypergeometric distribution in a runs test. His runs test looks at the probability of a particular number of success runs in a fixed sample size. The results of the sampling are recorded as either successes or failures and the number of runs is denoted R. The distribution of R is defined for both when R is an even number and when R is an odd number. The distribution is given below for both cases:

$$P(R = 2i) = \frac{2\binom{m-1}{i-1}\binom{n-1}{i-1}}{\binom{m+n}{m}}$$

$$P(R = 2i + 1) = \frac{\binom{m-1}{i}\binom{n-1}{i-1} + \binom{m-1}{i-1}\binom{n-1}{i}}{\binom{m+n}{n}}$$

where m is the number of successes and n is the number of failures.

Then, instead of focusing on the distribution of the number of runs, Guenther explores the connection between the distribution of the number of runs and the hypergeometric distribution. Guenther shows that the distribution of R can be written in the form of the hypergeometric distribution and the properties of the hypergeometric distribution can be applied to the distribution of the number of runs.

At this point we have given an overview of the hypergeometric distribution that incudes the origin of the word 'hypergeometric', derivation and elementary applications of the hypergeometric distribution, relationships that tie the hypergeometric distribution to other probability models, properties of the hypergeometric distribution, kindred hypergeometric distributions and statistical inference involving the hypergeometric distribution. As a result of examining literature, there were two articles that stood out. The first was an article by Walton (1986) on the number of observed classes from a subdivided population and the second was by Charlambides (2005) who looked at with replacement sampling from a subdivided population until a particular number of classes had been seen. We will extend the work of Walton to a sequential sampling method and

modify the work of Charlambides for without replacement sampling and extend it to a sequential sampling method as well.

2. DETERMINING A SURVIVING SUBPOPULATION USING SEQUENTIAL SAMPLING

Consider a population of size N, the members of which are divided into k classes. If sampling is conducted without replacement from the population, the question, "How many classes have been observed in this sample?" can be posed. This question is addressed by Walton (1986) in his article *The Number of Observed Classes From a Multiple Hypergeometric Distribution.* Walton built on the work of Emigh (1983) who answered the corresponding question when the sampling is done with replacement.

By understanding what Walton developed, we will extend his ideas to answer further questions regarding sampling without replacement from a population composed of k categories. For example, this idea could be thought of as the distribution of colors within a bag of candy or the distribution of colored balls in an urn. Suppose that a bag of chocolate candy has pieces that are blue, red, brown, yellow and orange. If a small bag has 4 blue, 6 red, 2 brown, 3 yellow and 5 orange, Walton's methods can be used to determine the probability that only red, yellow and orange are seen in a selection of size 8.

Similar to Walton's notation, consider a population that is made up of N^1 members and divided into k categories of sizes N_i where $i = 1,2, \dots, k$ and $\sum N_i = N^1$. We will use superscripts to denote the stages of sampling iteration, the meaning of which will become clear. Also, let m denote the number of observed classes in the sample. This

is actually the variable we seek to explore. Here, N^1 and k are both fixed values. However, the range of values of m depends on the sample size taken from the population. For example, if a population of size 20 were divided into 6 subcategories, it would be impossible for all 6 categories to be observed in only 4 draws.

If samples of size n are being taken from a population of size N^1, then the total number of ways this can be done is given by the combination, $\binom{N^1}{n}$. Also, because the sampling is being conducted without replacement, and the population is divided into distinct categories, the counts of the number of members per class has a multivariate hypergeometric distribution. Walton uses the notation $M(n, m, N)$ to denote the total number of potential samples of size n in which m classes are observed. Thus, simply stated, the distribution of the number of classes represented (X) in a sample of size n is given by:

$$P(X = m|N, n) = \frac{M(n,m,N)}{\binom{N}{n}}.$$

The question now becomes, how can $M(n, m, N)$ be calculated? Walton re-expresses $M(n, m, N)$ using a list denoted as C and a function he denotes as U. Using those, he is able to rewrite the above probability function as:

$$P(X = m|N, n) = \frac{\sum_{i=1}^{m}\binom{k-i}{k-m}(-1)^{m-i}U(n, i, N)}{\binom{N}{n}}$$

where $U(n, i, N) = \sum_{X \in C_i(N)}\binom{X}{n}$ and $C_i(N)$ lists all possible sums that are formed when taking i objects from N=[N_1 N_2 N_3 ... N_k]. Understanding the above expression for

$M(n, m, N)$, and most importantly the C and U functions, is most easily done by looking at an example.

To do so, consider a population of size $N = 10$ divided into $k = 4$ subcategories. Assume that the original distribution of N is given to be

$$N = [1 \quad 2 \quad 3 \quad 4].$$

The above notation is used to show the number of members in each of the $k = 4$ categories. That is, category one has 1 member, category two has 2, category three has 3 and category four has 4. Now, using what Walton defines to be the C function, C_i, $i = 1, \dots, k$, each of the C_i's can be calculated by adding all possible groups of size i to form a collection of sums. For example, C_1 takes all groups of size 1 and creates a new set comprised of their possible cardinalities. This is equivalent to the original population distribution. So $C_1 = [1, 2, 3, 4]$. Using the same logic, C_2 adds the cardinalities of all groups of size two from the original population to create a new set. In other words, C_2 gives the sum of the number of elements in all $\binom{4}{2}$ combinations of the original population distribution. By calculating the six sums, we see that $C_2 = [3, 4, 5, 5, 6, 7]$. These elements are a result of the sums $1 + 2, 1 + 3, 1 + 4, 2 + 3, 2 + 4$, and $3 + 4$. Similarly, C_3 can be calculated by adding all the cardinalities of the elements comprising the possible $\binom{4}{3}$ groups of 3 from the original population distribution. Calculating the four sums ($1 + 2 + 3, 1 + 2 + 4, 1 + 3 + 4$, and $2 + 3 + 4$) results in $C_3 = [6, 7, 8, 9]$. Similarly, there is only one possible way to add all the elements together to obtain C_4.

Thus, C_4 is simply the sum of the category sizes in N, the population size, and in this case, $C_4 = [10]$.

Now that each value of C_i has been calculated for the example, the U function can be used to create a table of values that will later be used in the calculation of $M(n, m, N)$. The U function sums combinatorial terms according to a particular value of C_i. In this example, $n = 1, \dots, 10$ while $i = 1, \dots, 4$. The following table consists of all the values of U for all combinations of n and i. To illustrate what is being calculated in the table below let's look at $U(3, 2, 10)$. Using the formula given by Walton,

$$\begin{aligned} U(3,2,10) &= \sum_{X \in C_2(N)} \binom{X}{3} \\ &= \binom{3}{3} + \binom{4}{3} + \binom{5}{3} + \binom{5}{3} + \binom{6}{3} + \binom{7}{3} \\ &= 1 + 4 + 10 + 10 + 20 + 35 \\ &= 80. \end{aligned}$$

Each of the forty calculations below were done in the same way to find the values of $U(n, i, N)$.

n	**i=1**	**i=2**	**i=3**	**i=4**
1	10	30	30	10
2	10	65	100	45
3	5	80	195	120
4	1	61	246	210
5	0	29	209	252
6	0	8	120	210
7	0	1	45	120
8	0	0	10	45
9	0	0	1	10
10	0	0	0	1

Table 2: U(n, i, N) for Original Population

The two above components, the C and U functions, can then be used to calculate the values of $M(n, m, N)$ needed to finally calculate the probability that a particular number of classes are observed in a sample of size n.

The values of $M(n, m, N)$ can also be summarized in a table like the one above. The values of $M(n, m, N)$ for this example are:

n	m=1	m=2	m=3	m=4
1	10	0	0	0
2	10	35	0	0
3	5	65	50	0
4	1	58	127	24
5	0	29	151	72
6	0	8	104	98
7	0	1	43	76
8	0	0	10	35
9	0	0	1	9
10	0	0	0	1

Table 3: M(n, m, N) for Original Population

Notice here that each of the rows in the table sum to $\binom{N}{n}$. Thus, for any given n, the values above are divided by $\binom{N}{n}$ to calculate the probability that a particular number of classes, 1 through 4, are observed in n draws.

From this point forward, the notation of Walton (1986) will still be used, but his investigation is being extended. Walton looked only at this process being iterated a single time. We will extend his work to an iterated process where a rule is put in place that governs the size of the sample taken in each iteration.

Now that Walton's notation has been explained, consider iterating a process of this nature until a single subgroup of the original population "survives" and each of the $k-1$ other subcategories "dies out" and has no representation. To explore this scenario, we establish a sampling rule. This sampling protocol regulates the number of draws taken in each iteration of the process. One possible way to define a sampling rule is to base it on percentages. In this way, the rule is defined to be that 100p% of the original population is collected for the first sample. This process is continued and the subsequent sample sizes are just 100p% of the previous sample each time. In cases where our sample size is a non-integer, we take the convention of rounding up to the nearest integer.

Using the example from above, where a population of size 10 is divided into 4 distinct subgroups, a selection rule can be established so that, in each iteration, 50% of the previous group is taken. Thus, the first sample taken from the population would be of size 5, because 50% applied to the population size of 10 results in 5 ($N^1 = 5$). Our question then becomes, what is the likelihood that there would only be one surviving

subcategory after each sampling iteration? In the case of this example, because the original population distribution is [1, 2, 3, 4], if 5 samples were taken in stage 1, there is no chance of observing only one class, or one subgroup being the "winner" or "survivor" after one stage of sequential sampling.

So, the question becomes, what is the chance that there is one remaining subcategory after *two* stages of sampling? For this example, stage 2 of the sampling would sample 50% of the 5 items taken during stage one. Because 50% of 5 comes out to be 2.5, the sample size for the second iteration of sampling would be rounded up to 3. That is, utilizing previously defined notation, $N^2 = 3$. In order to illustrate how calculations would be performed in this case, a simulation can be done to see which subcategories remained occupied after the first stage of sampling. Realize here, it is possible for each of the 4 subgroups to be represented after the first stage of sampling, but it is also possible that only 2 of the 4 subgroups are represented after the first stage of sampling. To simulate the sampling, a simulation using the *R* computing environment will be done in which the member of subgroup one is labeled as element 1, the members of subgroup two are labeled as elements 2 and 3, the members of subgroup three are labeled as elements 4, 5, and 6 and the members of subgroup four are labeled as elements 7, 8, 9, and 10.

Selecting a sample of 5 from the original population, labeled 1-10, resulted in the computer simulated sample [1, 4, 6, 7, 10] which means that *categories* 1, 3, and 4 are all still represented after the first stage of sampling. The question of how likely it is that 2, 3,

or 4 categories are represented after the first stage could also be posed. This question can be answered also by applying Walton's formula to the original sample. Because the chart for all possible values of $M(n, m, N)$, where $N = 10$, has already been constructed, calculating the probability that 2, 3 or 4 s are occupied after a sample of size 5 is taken from the population is simply putting the corresponding value of M into the formula

$$P(X = m|n, N) = \frac{M(n, m, N)}{\binom{N}{n}}.$$

Thus, the probability that when sampling 5 objects from the original 10, only 2 of the subcategories are still occupied is:

$$P(X = 2|5, 10) = \frac{M(5,2,10)}{\binom{10}{5}} = \frac{29}{252} \approx 0.115$$

The probability that 3 of the subcategories are still occupied, as in the *R* simulated example, is:

$$P(X = 3|5, 10) = \frac{M(5,3,10)}{\binom{10}{5}} = \frac{151}{252} \approx 0.599$$

And the probability that all 4 of the subcategories are still occupied is:

$$P(X = 4|5, 10) = \frac{M(5,4,10)}{\binom{10}{5}} = \frac{72}{252} \approx 0.286$$

Now, focusing on the R simulated example where subcategories 1, 3 and 4 are still occupied after the first stage of sampling, the question now becomes one of how likely it is that in the second stage of sampling only 1 subcategory remains occupied and can be declared the winner or survivor. Essentially, what has happened after the first

stage of sampling is that the population has been reduced by 50%, so the population size is now 5 ($N^2 = 5$) for calculation purposes. Similarly, the number of subgroups has been reduced from 4 to only 3.

The C and U functions calculated for the original population must now be calculated again with $N^2 = 5$ and $k = 3$. The number of elements in subgroup one is still 1 ($N_1 = 1$), the number in subgroup three is 2 ($N_3 = 2$) and the number in subgroup four is also 2 ($N_4 = 2$). The population can now be represented as

$$[1 \quad - \quad 2 \quad 2].$$

Thus, $C_1 = [1, 2, 2]$, $C_2 = [3, 3, 4]$, and $C_3 = [5]$. The values of U can once again be summarized in the following table of values:

n	i=1	i=2	i=3
1	5	10	5
2	2	12	10
3	0	6	10
4	0	1	5
5	0	0	1

Table 4: U(n, i, N) for Iteration Two

Thus, using the above values of U, a corresponding table of values for $M(n, \square, N)$ can be made:

n	m=1	m=2	m=3
1	5	0	0
2	2	8	0
3	0	6	4
4	0	1	4
5	0	0	1

Table 5: M(n, m, N) for Iteration Two

In order to find the probability that only 1 subcategory remains occupied after selecting 3 objects from the 5 remaining, the only value of M needed is $M(3,1,5) = 0$. This is because the remaining filled subcategories each have less than 3 members. Thus, there is no chance that a winner is chosen after 2 stages for this example.

Using R once again to simulate the sampling in stage two, three elements will be taken without replacement from the five elements left after stage one, [1, 4, 6, 7, 10]. The 3 elements in the sample above were [6, 7, 10] which means that subgroups 3 and 4 are both still represented. The probability that two subcategories remained occupied from this stage of sampling can be found from the above tables to be:

$$P(X = 2|3,5) = \frac{M(3,2,5)}{\binom{5}{3}} = \frac{6}{10} = \frac{3}{5} = 0.6.$$

At this point, a third stage of sampling would occur where 2 items are sampled from the three remaining. Now, only two subgroups remain and the population size has been reduced to three items only. Because subgroup 3 has one member and subgroup 4 has two members, $C_1 = [1,2]$ and $C_2 = [3]$. Thus, the values of U can be found to be:

n	i=1	i=2
1	3	3
2	1	3
3	0	1

Table 6: U(n, i, N) for Iteration Three

And the values of $M(n,m,N)$ can be found to be:

n	m=1	m=2
1	3	0
2	1	2
3	0	1

Table 7: M(n, m, N) for Iteration Three

The next round of sampling will select 2 items from the three left ($N^3 = 2$), and the probability that one subcategory remains is given to be

$$P(X = 1|2,3) = \frac{M(2,1,3)}{\binom{3}{2}} = \frac{1}{3} \approx 0.333.$$

Using *R* to simulate the drawing of two items from the remaining three, gives that elements 6 and 7 both remain and thus, subcategories 3 and 4 are still both occupied. This occurs with probability

$$P(X = 2|2,3) = \frac{M(2,2,3)}{\binom{3}{2}} = \frac{2}{3} \approx 0.667.$$

In the fourth and final sampling stage a final category will be chosen in a single draw from the two remaining subcategories ($N^4 = 1$). At this point, it can be clearly seen, that with two equally likely subcategories, the probability that category 3 is the ultimate "winner" ($P(X = 1|1,2)$) is $\frac{1}{2}$ and the probability that subcategory 4 is the ultimate "winner" ($P(X = 1|1,2)$) is also $\frac{1}{2}$. Using *R* to simulate the results, number 6 was drawn first and thus it was determined that subcategory 3 is the ultimate winner in this simulation.

At this point, we can calculate the probability that the sequential sampling were to turn out just as it did in the *R* simulation. Because each probability depends on the

population size, sample size and the number of observed classes, there are multiple ways to have subcategory 3 win in four sampling stages. Thus, only the probability that the subcategory 3 wins from the above iterations will be considered. When only considering one of the ways for subcategory 3 to win in four sampling stages, the probabilities calculated above can be multiplied to find the probability that things turned out exactly as *R* simulated. Thus,

$$P(cell\ 3\ wins | results\ above)$$

$$= P(X = 3|5, 10) * P(X = 2|3, 5) * P(X = 2|2, 3) * P(X = 1|1,2)$$

$$= 0.599 * 0.6 * 0.667 * 0.5$$

$$= .12.$$

Other questions of interest that are not within the scope of this thesis are questions such as, what is the probability that subcategory three is the winner without specifying the number of rounds required? Or, what is the probability that four stages are required? Or, only two stages are required? These questions, while directly related to the above, prove to be more complex to answer. In order to consider these questions, other questions such as, "how many possible ways are there for subcategory three to win?" must be considered. Because of the more complex nature, this is left for future investigations and is not covered at the present time.

The question now becomes, where can this be used? Though this concept is not widely used in literature, we believe that this idea can be applied in many different fields. There are many different processes, contests and elementary ideas such as the balls in

boxes and drawing of colored candy that can be represented by this type of process. With different selection rules in place, the applicability of this process becomes more evident. Some more specific applications that came to mind during this investigation were the selection of a particular Senator or House of Representative member to serve on a committee or something of similar nature. Also, with a sampling rule not based on percentages this could be applied to selecting student test scores at random to evaluate school performance.

After evaluating the work of Walton and extending it to a sequential sampling idea, we began to wonder what other questions can be asked in reference to subdivided populations and sequential sampling? This led us to the idea of sampling from a subdivided population *until at least one item was in each category*. This is similar to the relationship between the standard hypergeometric and negative hypergeometric in that in Walton's case, the sample size was fixed, and in this new idea, our sample size is a random variable. This new idea is explained in Chapter 3.

3. SEQUENTIAL SAMPLING TO ACHIEVE COMPLETE REPRESENTATION OF SUBPOPULATIONS

Charlambides (2005) presents a sequential random occupancy model for sampling with replacement from an infinite population. Charlambides does so by using the illustration of balls in urns, and this illustration will be carried through this section to describe the selection process. While Charlambides is concerned with sampling with replacement, we will modify his sequential random occupancy model in order to model the case of sampling without replacement. This will be done by looking first at the notation introduced by Charlambides, explaining it, and then presenting the modifications for sampling without replacement.

Once Charlambides' notation has been modified to the sampling without replacement case, the goal is to *iterate* this sampling scheme in order to have only one ball in each subcategory at the end of the process. This is the notion of complete representation of each subpopulation. In accomplishing this, the question "How many balls must be sampled in order to see at least one ball in each subcategory?" will be asked multiple times. This question is motivation to achieve our complete representation of the subcategories by iteratively sampling and achieving one ball in each subcategory.

3.1: Development and Modification of Charlambides Notation for Sampling Without Replacement

To modify the idea introduced by Charlambides, consider a population of size N,

divided into k subcategories of size $N_1, N_2, \dots, N_k$ where $\sum_{i=1}^{k} N_i = N^1$. Suppose the balls are selected from the population one at a time, *without replacement*. Let the total number of balls drawn from the urn at any one time be n. Now, let W_k denote the number of balls that must be drawn in order for all k urns to be occupied. Also, let M be the number of occupied urns after n draws from the population of size N. The probability $P(M = m) \equiv p_m(n, k)$ where $m = 1,2, \dots, k$ can be found by using the complement rule. Define A_i to be the event in which the i^{th} urn remains empty. Then, the probability that M urns are occupied is equivalent to the probability that $k - m$ of the events $A_i, i = 1, \dots, k$ occur, or that k-m urns remain empty after n draws.

Charlambides chooses to denote the events A_i with a double subscript where the first subscript indicates a particular subcategory and the second provides a count from 1 to r of the r events of concern. When sampling *with replacement*, Charlambides gives that the probability of selecting a particular ball on any particular draw as p_j. Then using the complement rule, Charlambides shows that

$$P(A_{i1}, A_{i2}, \dots, A_{ir}) = (1 - p_{i1} - p_{i2} - \cdots - p_{ir})^n = \left(p_{j1} + p_{j2} + \cdots + p_{j,k-r}\right)^n$$

where the j's just indicate the remaining k-r cells.

The above can now be modified to calculate the probability of r specific cells being selected when sampling *without replacement*. Just as Charlambides did when sampling with replacement, we will use the complement rule to give the probability that k-r urns remain empty after n draws. The fact that the sampling is being done without

replacement gives rise to the hypergeometric structure. Instead of using Charlambides' double subscripted notation, we will instead denote the number of elements in the $k-r$ filled cells as N_{f_i}, $i = 1, \dots, k-r$ and denote the number of elements in the r empty cells as N_{e_j} where $j = 1, \dots, r$. There are $\binom{N}{n}$ possible ways to select n items from a population of N, and more specifically $\binom{N_{e_1}+N_{e_2}+\cdots+N_{e_r}}{n}$ ways for the n items selected to come from a particular group of r empty cells. However, because the events A_i mean that a cell remains *empty* rather than filled, the complement rule can then be applied or we can instead look at the filled cells rather than the empty. Thus, the probability that $k-r$ particular cells are filled (r cells remain empty) when choosing items without replacement is given by

$$P\left(A_{e_1}, A_{e_2}, \dots, A_{e_r}\right) = \frac{\binom{N_{e_1}+N_{e_2}+\cdots+N_{e_r}}{n}}{\binom{N}{n}} = \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_{k-r}}}{n}}{\binom{N}{n}}.$$

Referring back to the work of Charlambides when sampling with replacement, a sum over all combinations of size $k-r$ can be defined as

$$S_{k,r} = \sum\left(p_{j1} + p_{j2} + \cdots + p_{j,k-r}\right)^n.$$

From the modification of $P\left(A_{e_1}, A_{e_2}, \dots, A_{e_r}\right)$ given above we can see that when sampling *without replacement* $S_{k,r}$ becomes

$$S_{k,r} = \sum \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_{k-r}}}{n}}{\binom{N}{n}}.$$

Notice, the expression for $S_{k,r}$ when sampling with replacement involves the power n. This power is no longer needed in our modification because the combinatorial term in the numerator incorporates the effect of sampling without replacement.

At this point, it is important to recall the inclusion-exclusion principle for the purpose of counting the number of elements in the union of a fixed number of sets. To begin, we sum all of the elements in the individual sets that comprise the union. From this sum, the cardinality of all the even-way intersections is subtracted (exclusion) while the cardinality of all odd-way intersections is added (inclusion). By applying the inclusion exclusion principle, Charlambides gives that the probability $P(M = m) \equiv p_m(n, k)$ can be given by the mass function, where $m = 1,2, \dots, k$,

$$p_m(n, k) = \sum_{r=0}^{m} (-1)^{m-r} \binom{k-r}{m-r} S_{k,k-r}.$$

Using the above, Charlambides shows that $P_m(n, k) \equiv P(M \leq m)$ can be given by the following where $m = 1,2, \dots, k$:

$$P_m(n, k) = \sum_{r=0}^{m} (-1)^{m-r} \binom{k-r-1}{k-m-1} S_{k,k-r}.$$

Similar to W_k, define W_m to be the number of balls that are sampled until $m \leq k$ urns are occupied. The probability distribution of W_m can be found by using the above result to see that

$$P_m(n, k) = P(M \leq m | S_k = n) = \sum_{r=0}^{m} (-1)^{m-r} \binom{k-r-1}{m-r-1} S_{k,k-r}$$

where $S_{k,k-r}$ is defined earlier, recalling that n is the number of balls sampled, k is the number of urns, and m indicates the number of occupied urns. Charlambides further shows that using $P_m(n,k)$, the probability that n draws are required in order to have m occupied urns (W_m) when sampling with replacement is

$$q_n(m,k) = P_{m-1}(n-1,k) - P_{m-1}(n,k) = \sum_{r=0}^{m-1} (-1)^{m-r-1} \binom{k-r-1}{m-r-1} T_{k,k-r}$$

where $T_{k,k-r}$ is given by

$$T_{k,k-r} = \sum \left[1 - \left(p_{j1} + p_{j2} + \cdots + p_{jr}\right)\right]\left(p_{j1} + p_{j2} + \cdots + p_{jr}\right)^{n-1},$$

where the sum is taken over all combinations of size r.

Recall that the difference of two binomial probabilities can at times result in a negative binomial probability. To aid in understanding the equation $q_n(m,k) = P_{m-1}(n-1,k) - P_{m-1}(n,k)$, consider the question, "What is the chance that four flips of a coin are required to see three heads for the first time?" We will use the above definition of q to answer this question. For our illustration, $m = 3$ for the three heads that are desired, so $m - 1 = 2$ and $n = 4$ for the number of flips required, thus $n - 1 = 3$. Thus, $P_{m-1}(n-1,k)$ represents the probability that there are two or less heads in three trials. There are eight possible outcomes for three coin flips: {HHH, HHT, HTH, HTT, TTT, TTH, THT, THH}. We can see that seven out of those eight {HHT, HTH, HTT, TTT, TTH, THT, THH} indicate that there were two or less heads in three trials. So, $P_{m-1}(n-1,k) = P_2(3,k) = \frac{7}{8}$. Now, we must find the chance that there are two or less

heads in four trials to calculate $P_{m-1}(n,k) = P_2(4,k)$. There are sixteen possible outcomes for flipping four coins given by {HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, THHH, THHT, HTTH, HTTT, TTTH, TTTT, TTHH, TTHT, THTH, THTT}. We can see that eleven out of those sixteen {HHTT, HTHT, THHT, HTTH, HTTT, TTTH, TTTT, TTHH, TTHT, THTH, THTT} have at most two heads in four trials. Thus $P_{m-1}(n,k)$ for our problem is given by $P_{m-1}(n,k) = P_2(4,k) = \frac{11}{16}$. Subtracting these two probabilities results in $q_n(m,k) = q_4(3,k) = \frac{7}{8} - \frac{11}{16} = \frac{3}{16}$. To show this is equivalent to a negative binomial probability as stated above, consider the probability that we have two heads in the first three flips given by: $\binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = \frac{3}{8}$ and the probability that we flip a head on the fourth flip, in order to have three heads for the first time in exactly four flips, given by $\frac{1}{2}$. Multiplying the two together gives the negative binomial probability that it takes exactly four flips to result in three heads for the first time, and the probability is $\frac{3}{16}$ which is equivalent to that found by using the definition of the q function provided earlier.

In order to modify the T function given earlier for sampling without replacement, we will focus on the fact just illustrated that

$$q_n(m,k) = P_{m-1}(n-1,k) - P_{m-1}(n,k).$$

We will use the definitions of $P_m(n,k)$ and $S_{k,k-r}$ for without replacement sampling given earlier, to rewrite the above subtraction. This gives

$$q_n(m,k) = \sum_{r=0}^{m-1} (-1)^{m-1-r} \binom{k-r-1}{k-m} \left[\sum \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_r}}{n-1}}{\binom{N}{n-1}} \right]$$

$$- \sum_{r=0}^{m-1} (-1)^{m-1-r} \binom{k-r-1}{k-m} \left[\sum \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_r}}{n}}{\binom{N}{n}} \right],$$

where the sums inside the brackets are taken over all combinations of size r.

Factoring out the common terms, we can deal now with simplifying

$$\frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_r}}{n-1}}{\binom{N}{n-1}} - \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_r}}{n}}{\binom{N}{n}}.$$

By writing out the combinatorial terms and multiplying by "convenient one" terms, the above can be shown to be

$$\frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_r}}{n-1}}{\binom{N}{n-1}} * \frac{N-(N_{f_1}+N_{f_2}+\cdots+N_{f_r})}{N-n+1}.$$

Putting it all together, we have shown that for *sampling without replacement*, $q_n(\square, k) =$

$$\sum_{r=0}^{m-1} (-1)^{m-r-1} \binom{k-r-1}{k-m} \left[\sum \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_r}}{n-1}}{\binom{N}{n-1}} * \frac{N-(N_{f_1}+N_{f_2}+\cdots+N_{f_r})}{N-n+1} \right].$$

So, by definition of $q_n(m,k)$ given above we can see that in the case of *sampling without replacement*,

$$T_{k,k-r} = \sum \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{f_r}}{n-1}}{\binom{N}{n-1}} * \frac{N-(N_{f_1}+N_{f_2}+\cdots+N_{f_r})}{N-n+1}.$$

The above can now be used to calculate the probability that we require n draws to occupy m urns. For our specific problem, the idea is focused on finding the probability that n draws are required to occupy *all* of the k urns.

Recall the T function without replacement defined earlier,

$$T_{k,k-r} = \sum\left[1 - \left(p_{j1} + p_{j2} + \cdots + p_{jr}\right)\right]\left(p_{j1} + p_{j2} + \cdots + p_{jr}\right)^{n-1}.$$

Notice that the T function when sampling with replacement has geometric structure. This relates to the geometric mass function because the idea is waiting on the first success after a particular number of successes have been observed. Recall, the geometric mass function is a particular case of the negative binomial distribution. Thus, extending this idea to sampling without replacement, the structure of the T function, given below, is now negative hypergeometric.

All of the prior development allows us to use the equation

$$q_n(m,k) = \sum_{r=0}^{m-1} (-1)^{m-r-1} \binom{k-r-1}{m-r-1} T_{k,k-r},$$

where

$$T_{k,k-r} = \sum \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{fr}}{n-1}}{\binom{N}{n-1}} * \frac{N - (N_{f_1} + N_{f_2} + \cdots + N_{f_r})}{N-n+1}.$$

These equations will be used to answer the question, "When sampling without replacement, what is the probability that all k urns are occupied after n draws?" At this point, the sequential sampling can be done and the above question can be asked at each

stage until only one ball remains in each urn. This is illustrated in the following simulation using the *R* computing environment.

3.2: Simulated Example

In order to perform calculations in context, *R* was used to simulate the drawing of balls from a subdivided population. The original population for the example consisted of $N^1 = 20$ members divided among 6 distinct categories. To visually represent this, consider the following graphic:

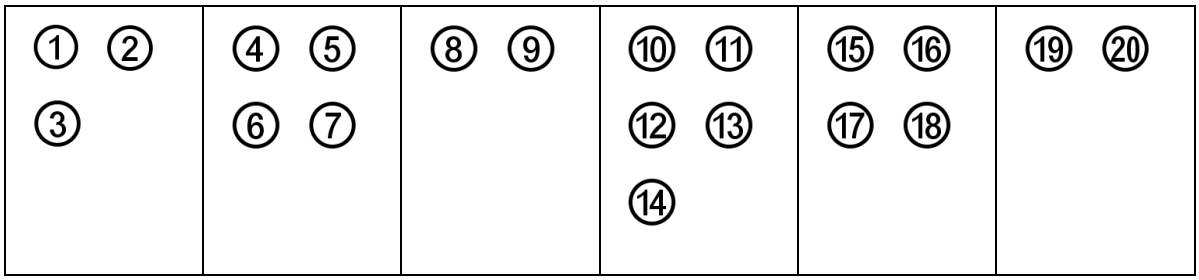

Figure 1: Original Population Distribution

In cell one there are three members, cell two has four, cell three has two, cell four has five, cell five has four and cell six has two members. Thus, we have $N_1^1 = 3,\ N_2^1 = 4, N_3^1 = 2, N_4^1 = 5, N_5^1 = 4$ and $N_6^1 = 2$. This gives the original distribution of the population. Say now that balls are drawn one at a time without replacement. The problem we choose to explore is sequentially sampling in stages until there is only one member remaining in each cell. At this point, all of Charlambides notation has been explained and modified to fit our current situation of sampling without replacement. The modified equations developed in Section 3.1 will be used to calculate the probabilities for this example in Section 3.3.

We will use *R* to generate random numbers and label the balls so that the balls in cell one are labeled 1-3, the balls in cell two are labeled 4-7, the balls in cell three are labeled 8-9, the balls in cell four are labeled 10-14, the balls in cell five are labeled 15-18, and the balls in cell six are labeled 19-20. The random number sequences determine which balls are designated as the representatives for the next round of sampling. *R* will generate the random number sequence for the first round of sampling by giving a random permutation of the numbers 1 through 20. Once this sequence is generated, the numbers in the sequence tell us which balls to sequentially place in their designated cells. Once there is at least one ball in each cell the remaining numbers in the sequence can be disregarded. For instance, if R generates the sequence {6, 12, 1, 16, 15, 2, 3, 18, 10, 17, 14, 7, 20, 9, 19, 8, 4, 5, 13, 11}, call this "Round 1", or the first iteration, of the process. Then using this sequence, balls 19, 8, 4, 5, 13, and 11 can be "thrown out" because it took only 14 draws to get at least one ball in each cell. Once ball number 9 was placed in it's corresponding cell, there was at least one ball in each of the six cells, so the sampling would stop after 14 draws. Thus, for the purpose of calculations, n=14 in order for all six cells to be filled. The resulting cell distribution would be:

① ② ③	⑥ ⑦	⑨	⑩ ⑫ ⑭	⑮ ⑯ ⑰ ⑱	⑳

Figure 2: Distribution after Iteration One

Now for the second round of sampling, the above cell distribution can now be viewed as a "population" itself, the size of which is denoted $N^2 = 14$. At this point *R* can be used to generate a random sequence of the remaining numbers representing the balls in this second round of sampling which are: {6, 12, 1, 16, 15, 2, 3, 18, 10, 17, 14, 7, 20, 9}. The resulting sequence from R is {17, 9, 2, 7, 12, 14, 3, 1, 6, 10, 16, 20, 15, 18}. Proceeding as before, after ball number 20 is drawn, each cell has at least one member, so balls 15 and 18 can be eliminated. This means that modifying the population to be only the members left after Round 1, it took 12 draws to see at least one member in each cell from our population of size 14. This leaves the following in each of the six cells:

① ② ③	⑥ ⑦	⑨	⑩ ⑫ ⑭	⑯ ⑰	⑳

Figure 3: Distribution after Iteration Two

Thus, it can be seen that $N^3 = 12$. Similarly, repeating the process once more, R generates the following sequence from the remaining balls above: {6, 7, 9, 17, 20, 2, 16, 3, 1, 12, 10, 14}. It is easily seen that after modifying the population to be only the members left after round two of sampling, it only took n=10 draws to get at least one member in each cell. In this sequence 10 and 14 can be eliminated leaving the following distribution of balls:

① ② ③	⑥ ⑦	⑨	⑫	⑯ ⑰	⑳

Figure 4: Distribution after Iteration Three

Now, $N^4 = 10$ and the next sequence (excluding now 10 and 14) generated from R is: {6, 16, 12, 17, 2, 20, 9, 7, 3, 1}, that means balls 6, 16, 12, 17, 2, 20 and 9 are placed into their corresponding cells and 7, 3, and 1 can be eliminated. We can see that in this case n=7 and the remaining balls are distributed as follows:

②	⑥	⑨	⑫	⑯ ⑰	⑳

Figure 5: Distribution after Iteration 4, 5, and 6

The next three sequences generated from *R* result in the same distribution as in Round 4. That is, $N^5 = N^6 = N^7 = 7$. On the 8th round of sampling, ball 17 is eliminated leaving only {2, 6, 9, 12, 16, 20} as the "winning" representations from each subcategory.

3.3: Probability Calculations

This example from Section 3.2 shows that for this particular case it took 8 rounds of sampling to achieve the situation where there was exactly one ball in each of the six cells. The question is then, what is the probability that this would have been the result? Or we could ask questions such as what is the probability that the desired result would have been achieved on the first round of draws?

This question, along with others, can be answered by referring to the modifications of Charlambides' work given in Section 3.1. In order to find the probability that all k=6 cells, in the case of the example, are occupied in a particular number of trials, n, we can use the equation for $q_n(m,k)$, where both m and k are given to be 6. It is important to note that while the initial population size is given to be 20, each time the sampling is done, members of the initial population are being "thrown out" once the desired "at least one in each" distribution is drawn/simulated.

Using the above example to illustrate the calculations, it must be carefully noted which iteration of the sampling is being conducted. The example is one of sequential sampling where the sampling scheme is the same at each stage, namely to sample until at least one member of each category is seen. Since the desire is to have *a* representative from each subcategory, these extra members are not needed.

Looking at the first iteration of the process, it can be seen that 14 draws were required to see at least one member in each cell. The probability that all 6 cells are occupied after 14 draws can be found using the equation for q given in Section 3.1 where T has been modified for sampling without replacement,

$$q_n(m,k) = \sum_{r=0}^{m-1} (-1)^{m-r-1} \binom{k-r-1}{m-r-1} T_{k,k-r} = T_{6,1} - T_{6,2} + T_{6,3} - T_{6,4} + T_{6,5} - T_{6,6}.$$

In this case, $m = k = 6$, and the combinatorial term in the equation reduces to 1. Thus the equation of interest for Round 1 is:

$$q_{14}(6,6) = \sum_{r=0}^{5} (-1)^{6-r-1}\, T_{6,6-r}.$$

The probability that all 6 cells are occupied after 14 draws can be found using the formula established in Section 3.1 for T:

$$T_{k,k-r} = \sum \frac{\binom{N_{f_1}+N_{f_2}+\cdots+N_{fr}}{n-1}}{\binom{N}{n-1}} * \frac{N-(N_{f_1}+N_{f_2}+\cdots+N_{f_r})}{N-n+1}.$$

The resulting calculations for $q_{14}(6,6)$ are as follows:

$q_{14}(6,6)$

$$= (-1)^5 \frac{\binom{0}{13}20}{\binom{20}{13}7} + (-1)^4 \frac{2\binom{2}{13}18 + \binom{3}{13}17 + 2\binom{4}{13}16 + \binom{5}{13}15}{\binom{20}{13}7}$$

$$+ (-1)^3 \frac{\binom{4}{13}16 + 2\binom{5}{13}15 + 4\binom{6}{13}14 + 4\binom{7}{13}13 + 2\binom{8}{13}12 + 2\binom{9}{13}11}{\binom{20}{13}7}$$

$$+ (-1)^2 \frac{\binom{7}{13}13 + 2\binom{8}{13}12 + 5\binom{9}{13}11 + 4\binom{10}{13}10 + 5\binom{11}{13}9 + 2\binom{12}{13}8 + \binom{13}{13}7}{\binom{20}{13}7}$$

$$+ (-1)^1 \frac{2\binom{11}{13}9 + 2\binom{12}{13}8 + 4\binom{13}{13}7 + 4\binom{14}{13}6 + 2\binom{15}{13}5 + \binom{16}{13}}{\binom{20}{13}7} 4$$

$$+ (-1)^0 \frac{\binom{15}{13}5 + 2\binom{16}{13}4 + \binom{17}{13}3 + 2\binom{18}{13}2}{\binom{20}{13}7}$$

$$\approx 0.0788183694531$$

Notice that many of the terms can be eliminated because the combinatorial terms in the numerator result in a value of zero, since we have $\binom{a}{b} = 0$ if $b > a$. From this point forward, such terms will not be shown in the calculations.

The same steps can be taken to find the probability that at any stage a particular number of draws is required to fill all 6 cells. Moving on to Round 2 of the process, 12 draws were required from the remaining 14 from iteration 1 to see an occupant in each of the 6 cells. This means $n = 12, N^2 = 14$ such that the distribution of the population is [3,2,1,3,4,1]. In order to calculate the current probability, the equation

$$q_{12}(6,6) = \sum_{r=0}^{5} (-1)^{6-r-1} T_{6,6-r}$$

will be used once again. Because the cell composition is different now, along with the population size, the calculations below will result in the probability that all 6 cells will be filled after 12 draws from a population of size 14.

$$q_{12}(6,6) = (-1)^1 \frac{\binom{12}{11}2 + 2\binom{11}{11}3}{\binom{14}{11}3} + (-1)^0 \frac{2\binom{11}{11}3 + \binom{12}{11}2 + 2\binom{13}{11}}{\binom{14}{11}3}$$

$$= \frac{1}{7} \approx 0.142857142857$$

Continuing through each iteration, for iteration 3, $n = 10$ and $N^3 = 12$ where the third population is represented as [3,2,1,3,2,1]. Thus, the corresponding probability at this stage is given by:

$$q_{10}(6,6) = (-1)^1 \frac{4\binom{9}{9}3 + \binom{10}{9}}{\binom{12}{9}3} + (-1)^0 \frac{2\binom{9}{9}3 + 2\binom{10}{9}2 + 2\binom{11}{9}}{\binom{12}{9}3}$$

$$\approx 0.187878787879$$

For iteration 4, $n = 7$ and $N^4 = 10$ and the fourth population is represented by [3,2,1,1,2,1] so that

$$q_7(6,6) = (-1)^2 \frac{\binom{7}{6}3 + 6\binom{6}{6}4}{\binom{10}{6}4} + (-1)^1 \frac{3\binom{8}{6}2 + 6\binom{7}{6}3 + 4\binom{6}{6}4}{\binom{10}{6}4}$$

$$+ (-1)^0 \frac{3\binom{9}{6} + 2\binom{8}{6}2 + \binom{7}{6}3}{\binom{10}{6}4}$$

$$= \frac{1}{7} \approx 0.142857142857.$$

Iterations 5,6, and 7 maintained a population size of 7 with cell distribution of [1,1,1,1,2,1]. The probability that this happens just once is found by

$$q_7(6,6) = (-1)^4 \frac{5\binom{6}{6}}{\binom{7}{6}1}$$

$$= \frac{5}{7} \approx 0.71428571.$$

Lastly, for the final iteration, $N^8 = 7$ where the distribution is the same as above, but the probability that in 6 draws all the cells are filled is given by

$$q_6(6,6) = (-1)^1 \frac{10\binom{5}{5}2}{\binom{7}{5}2} + (-1)^0 \frac{\binom{5}{5}2 + 5\binom{6}{5}}{\binom{7}{5}2}$$

$$= \frac{2}{7} \approx 0.285714285714.$$

We can see that, in this particular example, eight rounds of sampling were required to reach our desired end of having one ball in each cell. However, this is just one of the many ways that this result could have been achieved. It is possible that we could have reached our desired result in just one round of sampling if each of the first six balls corresponded to different subcategories. It is also possible that this process could have

taken more than eight rounds to achieve one ball in each cell. We can see that in our example, the same balls were drawn for three rounds in a row. This could have continued for any number of rounds or happened at any stage of the process.

Similar to the survival cell problem discussed in Chapter 2, to calculate the probability that a particular ball would be the winner of a subcategory or the probability that the sampling ended after a particular number of stages is beyond the scope of this thesis. These are very relevant questions, but because each calculation involves subgroup sizes, population size and sample size the answer to those questions become very complex very quickly.

Though we did not find this concept of sequentially sampling until complete subcategory representation anywhere in previously published literature, we feel that this idea can be applied to various real world situations. For instance, thinking of the subpopulations as states, countries or ethnicities leads to the idea of choosing a single member to serve as a representative of the entire subpopulation.

4. SUMMARY AND CONCLUSIONS

Through this investigation, we have looked at not only an overview of the hypergeometric distribution, but extensions of the hypergeometric distribution to sequential sampling with the purpose of having one single subcategory survive or to achieve complete representation of subcategories within a population.

Chapter One aims to cover some of the main features, properties and facts about the hypergeometric distribution as well as lesser known, lesser utilized properties and uses of the distribution. This was done by sorting literary results into the following sections: Origin of the Word 'Hypergeometric', Derivation and Elementary Applications, Relationships to Other Probability Models, Properties of the Hypergeometric Distribution, Kindred Hypergeometric Distributions and Statistical Inference. Through this we were able to find and become interested in Walton's (1986) idea of sampling from a subdivided population.

In Chapter Two, we extend Walton's idea of finding the observed number of subcategories within a subdivided population. Walton's notation and original problem are explained and we then extend his idea to a sequential sampling scheme. Instead of desiring the probability that a particular number of subcategories are occupied after a particular number of draws, we want to sequentially sample until there is only one "surviving" subpopulation. We do so by employing a sampling rule and the methods used

by Walton at each stage of sampling. Though the sampling rule we chose to explore was one based on percentages of the previous population, the sampling rule could be a function of the subcategory sizes or adapted to fit the desired number of samples to be taken at each stage.

Walton (1986) based his study off of the work done by Emigh (1983) who looked at finding the probability that m out of k classes were observed when sampling with replacement. Walton was able to use Emigh's work to create a model for without replacement sampling. Together their work covers finding probabilities for occupied categories when the sample size is fixed.

Charlambides (2005), however, looked at the "negative binomial" case where the samples are being taken with replacement until m categories have been observed. We noticed that there appeared to be a gap in the research covering the "negative hypergeometric" case. Thus, we modified Charlambides' work to model sampling without replacement until m categories were observed. We then extended this idea to a sequential sampling scheme in order to achieve only one member left in each category.

The following table can be seen as a summary both of the work done in this thesis and the work covered in earlier research.

	Trials Fixed	Success Fixed/Trials Random
With Replacement	Emigh (1983)	Charlambides (2005)
Without Replacement	Walton (1986) Busbee (2017)	Busbee (2017)

We have extended the ideas developed by Walton to sequential sampling. Also, we have modified the work of Charlambides to model sampling without replacement until a particular number of classes have been observed, then once again extended this idea to sequential sampling.

Bibliography

Alvo, M., & Cabilio, P. (2000). Calculation of Hypergeometric Probabilities Using Chebyshev Polynomials. *The American Statistician*, *54*(2), 141–144. https://doi.org/10.2307/2686033

Bardsley, W. E. (2016). Note on the hypergeometric distribution as an invalidation test for binary forecasts. *Stochastic Environmental Research and Risk Assessment*, *30*(3), 1059–1061. https://doi.org/10.1007/s00477-015-1071-z

Bardwell, G. E., & Crow, E. L. (1964). A Two-Parameter Family of Hyper-Poisson Distributions. *Journal of the American Statistical Association*, *59*(305), 133–141. https://doi.org/10.2307/2282864

Barnier, W., & Jantosciak, J. (n.d.). Duality and Symmetry in the Hypergeometric Distribution | Mathematical Association of America. Retrieved February 1, 2017, from http://www.maa.org/programs/faculty-and-departments/classroom-capsules-and-notes/duality-and-symmetry-in-the-hypergeometric-distribution

Bol'shev, L. (1964). Distributions Related to the Hypergeometric Distribution. *Theory of Probability & Its Applications*, *9*(4), 619–624. https://doi.org/10.1137/1109083

Bol'shev, L. N., & Seckler, Bernard. (n.d.). CIS - Distributions related to the hypergeometric distribution. Retrieved February 7, 2017, from https://statindex.org/articles/2696

Bosch, A. j. (1963). The Pólya distribution. *Statistica Neerlandica*, *17*(3), 201–213. https://doi.org/10.1111/j.1467-9574.1963.tb01040.x

Bowman, K. O., Kastenbaum, M. A., & Shenton, L. R. (1992). The negative hypergeometric distribution and estimation by moments. *Communications in Statistics - Simulation and Computation*, *21*(2), 301–332. https://doi.org/10.1080/03610919208813021

Broca, D. S. (2008). Derivation of the Hypergeometric Distribution: An Alternative Reasoning Approach. *International Journal of Mathematical Education in Science and Technology*, *39*(2), 269–273.

Brunk, H. D., Holstein, J. E., & Williams, F. (1968). A Comparison of Binomial Approximations to the Hypergeometric Distribution. *The American Statistician*, *22*(1), 24–26. https://doi.org/10.2307/2681878

Chae, K. C. (1993). Presenting the negative hypergeometric distribution to the introductory statistics courses. *International Journal of Mathematical Education in Science and Technology*, *24*(4), 523–526. https://doi.org/10.1080/0020739930240403

Charlambides, C.A. (2005). *Combinatorial Methods in Discrete Distributions*. Wiley.

Chesson, J. (1976). A Non-Central Multivariate Hypergeometric Distribution Arising from Biased Sampling with Application to Selective Predation. *Journal of Applied Probability*, *13*(4), 795–797. https://doi.org/10.2307/3212535

Childs, A., & Balakrishnan, N. (2000). Some approximations to the multivariate hypergeometric distribution with applications to hypothesis testing. *Computational Statistics & Data Analysis*, *35*(2), 137–154. https://doi.org/10.1016/S0167-9473(00)00007-4

Chvátal, V. (1979). The tail of the hypergeometric distribution. *Discrete Mathematics*, *25*(3), 285–287. https://doi.org/10.1016/0012-365X(79)90084-0

Cooper, John C. B. (2016). The Hypergeometric Distribution: an Application from Poker. Retrieved February 1, 2017, from http://connection.ebscohost.com/c/articles/112822000/hypergeometric-distribution-application-from-poker

Dacey, M. F. (1969). A Hypergeometric Family of Discrete Probability Distributions: Properties and Applications to Location Models. *Geographical Analysis*, *1*(3), 283–317. https://doi.org/10.1111/j.1538-4632.1969.tb00625.x

Davidson, R. R., & Johnson, B. R. (n.d.). Interchanging Parameters of the Hypergeometric Distribution | Mathematical Association of America. Retrieved February 7, 2017, from http://maa.org/programs/faculty-and-departments/classroom-capsules-and-notes/interchanging-parameters-of-the-hypergeometric-distribution

Davies, O. L. (1933). On Asymptotic Formulae for the Hypergeometric Series: I. Hypergeometric Series in Which the Fourth Element, x, is Unity. *Biometrika*, *25*(3/4), 295–322. https://doi.org/10.2307/2332287

D'Elia, A. (2003). Modelling ranks using the inverse hypergeometric distribution. *Statistical Modelling*, *3*(1), 65–78. https://doi.org/10.1191/1471082X03st047oa

Eisinga, R., & Pelzer, B. (2011). Saddlepoint approximations to the mean and variance of the extended hypergeometric distribution. *Statistica Neerlandica*, *65*(1), 22–31. https://doi.org/10.1111/j.1467-9574.2010.00468.x

Emigh, T. H. (1983). On the Number of Observed Classes From a Multinomial Distribution. *Biometrics*.

Fog, A. (2008). Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions. *Communications in Statistics - Simulation and Computation*, *37*(2), 241–257. https://doi.org/10.1080/03610910701790236

Fury, W., Batliwalla, F., Gregersen, P. K., & Li, W. (2006). Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. *Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference*, *1*, 5531–5534. https://doi.org/10.1109/IEMBS.2006.260828

Gart, J. J. (1963). A Median Test with Sequential Application. *Biometrika*, *50*(1/2), 55–62. https://doi.org/10.2307/2333746

Godbole, A. P. (1990). On hypergeometric and related distributions of order k. *Communications in Statistics - Theory and Methods*, *19*(4), 1291–1301. https://doi.org/10.1080/03610929008830262

Govindarajulu, Z. (1964). The First Two Moments of the Reciprocal of the Positive Hypergeometric Variable. *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, *26*(3/4), 217–236.

Guenther, W. C. (1978). Some Remarks on the Runs Test and the Use of the Hypergeometric Distribution. *The American Statistician*, *32*(2), 71–73. https://doi.org/10.1080/00031305.1978.10479256

Gupta, A. K., & Nagar, D. K. (2012). MATRIX-VARIATE GAUSS HYPERGEOMETRIC DISTRIBUTION. *Journal of the Australian Mathematical Society*, *92*(3), 335–355. https://doi.org/10.1017/S1446788712000353

Hald, A. (1960). The Compound Hypergeometric Distribution and a System of Single Sampling Inspection Plans Based on Prior Distributions and Costs. *Technometrics*, *2*(3), 275–340. https://doi.org/10.2307/1266247

Harkness, W. L. (1965). Properties of the Extended Hypergeometric Distribution. *The Annals of Mathematical Statistics*, *36*(3), 938–945.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014). *Nonparametric Statistical Methods* (3rd ed.). Wiley.

Hu, D. P., Cui, Y. Q., & Yin, A. H. (2013). An Improved Negative Binomial Approximation for Negative Hypergeometric Distribution. *Applied Mechanics and Materials*, *427–429*, 2549–2553. https://doi.org/10.4028/www.scientific.net/AMM.427-429.2549

Hush, D., & Scovel, C. (2005). Concentration of the hypergeometric distribution. *Statistics & Probability Letters*, *75*(2), 127–132. https://doi.org/10.1016/j.spl.2005.05.019

Irwin, J. O. (1954). A Distribution Arising in the Study of Infectious Diseases. *Biometrika*, *41*(1/2), 266–268. https://doi.org/10.2307/2333023

Irwin, J. O. (1975). The Generalized Waring Distribution. Part I. *Journal of the Royal Statistical Society. Series A (General)*, *138*(1), 18–31. https://doi.org/10.2307/2345247

Joarder, A. H. (2012). Some Instructional Issues in Hypergeometric Distribution. *Pakistan Journal of Statistics and Operation Research*, *8*(3), 669–684. https://doi.org/10.18187/pjsor.v8i3.536

Joarder, A. H., & Al-Sabah, W. S. (2007). Probability issues in without replacement sampling. *International Journal of Mathematical Education in Science and Technology*, *38*(6), 823–831. https://doi.org/10.1080/00207390701228385

Johnson, N. L., Kotz, S., & Kemp, A. W. (1992). *Univariate Discrete Distributions* (Second Edition). John Wiley and Sons.

Kemp, A. W. (2005). Steady-state Markov chain models for certain q-confluent hypergeometric distributions. *Journal of Statistical Planning and Inference*, *135*(1), 107–120. https://doi.org/10.1016/j.jspi.2005.02.009

Kemp, C. D., & Kemp, A. W. (1956). Generalized Hypergeometric Distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *18*(2), 202–211.

Kerov, S. V. (2005). Multidimensional Hypergeometric Distribution and Characters of the Unitary Group. *Journal of Mathematical Sciences*, *129*(2), 3697–3729. https://doi.org/10.1007/s10958-005-0309-6

Kumar, C. S. (2002). Extended generalized Hypergeometric probability distributions. *Statistics & Probability Letters*, *59*(1), 1–7. https://doi.org/10.1016/S0167-7152(02)00102-5

Lahiri, S. N., & Chatterjee, A. (2007). A Berry-Esseen Theorem for Hypergeometric Probabilities under Minimal Conditions. *Proceedings of the American Mathematical Society*, *135*(5), 1535–1545.

Lessing, R. (1973). The Teacher's Corner: An Alternative Expression for the Hypergeometric Moment Generating Function. *The American Statistician*, *27*(3), 115–115. https://doi.org/10.1080/00031305.1973.10479008

Liao, J. (1992). An Algorithm for the Mean and Variance of the Noncentral Hypergeometric Distribution. *Biometrics*, *48*(3), 889–892. https://doi.org/10.2307/2532354

Liao, J. G., & Rosen, O. (2001). Fast and Stable Algorithms for Computing and Sampling from the Noncentral Hypergeometric Distribution. *The American Statistician*, *55*(4), 366–369.

Liao, J., Munoz, A., & Rosner, B. (1987). On Cornfield's Mean and Variance for the Hypergeometric Distribution. *Biometrics*, *43*(3), 727–727.

Lieberman, G. J., & Owen, D. B. (1961). *Tables of the Hypergeometric Probability Distribution*. Stanford University Press.

Ling, R. F., & Pratt, J. W. (1984). The Accuracy of Peizer Approximations to the Hypergeometric Distribution, with Comparisons to Some Other Approximations. *Journal of the American Statistical Association*, *79*(385), 49–60. https://doi.org/10.2307/2288333

Marshall, A. W., & Olkin, I. (1985). A Family of Bivariate Distributions Generated by the Bivariate Bernoulli Distribution. *Journal of the American Statistical Association*, *80*(390), 332–338. https://doi.org/10.1080/01621459.1985.10478116

Mathai, A. M., & Saxena, R. K. (1967). On a generalized hypergeometric distribution. *Metrika*, *11*(1), 127–132. https://doi.org/10.1007/BF02613583

Mehta, C. R., & Patel, N. R. (1983). A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables. *Journal of the American Statistical Association*, *78*(382), 427–434. https://doi.org/10.1080/01621459.1983.10477989

Miller, G. K. (2006). *Probability: Modeling and Applications to Random Processes* (1 edition). Hoboken, NJ: Wiley-Interscience.

Miller, G. K., & Fridell, S. L. (2007). A Forgotten Discrete Distribution? Reviving the Negative Hypergeometric Model. *The American Statistician*, *61*(4), 347–350.

Nagar, D. K., & Sepúlveda-Murillo, F. H. (2009). Multivariate Generalization of the Confluent Hypergeometric Function Kind 1 Distribution. *International Journal of Mathematics and Mathematical Sciences*, *2008*, e152808. https://doi.org/10.1155/2008/152808

Nicholson, W. L. (1956). On the Normal Approximation to the Hypergeometric Distribution. *The Annals of Mathematical Statistics*, *27*(2), 471–483.

Ollero, J., & Ramos, H. M. (1995). Description of a subfamily of the discrete pearson system as generalized-binomial distributions. *Journal of the Italian Statistical Society*, *4*(2), 235–249. https://doi.org/10.1007/BF02589104

Plachky, D. (2003). Relationships Between the Negative Binomial and Negative Hypergeometric Distributions, with Applications to Testing and Estimation. *American*

Journal of Mathematical and Management Sciences, *23*(1–2), 1–6. https://doi.org/10.1080/01966324.2003.10737601

Rodríguez-Avi, J., Conde-Sánchez, A., Sáez-Castillo, A. J., & Olmo-Jiménez, M. J. (2007). Gaussian Hypergeometric Probability Distributions for Fitting Discrete Data. *Communications in Statistics - Theory and Methods*, *36*(3), 453–463. https://doi.org/10.1080/03610920601001733

Sandiford, P. J. (1960). A New Binomial Approximation for Use in Sampling from Finite Populations. *Journal of the American Statistical Association*, *55*(292), 718–722. https://doi.org/10.2307/2281594

Sathakathulla, A. A., & Murthy, B. N. (2012). Single, Double and Multiple Sampling Plans: Hypergeometric Distribution. *Journal of Interdisciplinary Mathematics*, *15*(4–5), 275–338. https://doi.org/10.1080/09720502.2012.10700800

Schuster, E. F., & Sype, W. R. (1987). On the negative hypergeometric distribution. *International Journal of Mathematical Education in Science and Technology*, *18*(3), 453–459. https://doi.org/10.1080/0020739870180316

Shuster, J. J. (1991). The Statistician in a Reverse Cocaine Sting. *The American Statistician*, *45*(2), 123–124. https://doi.org/10.1080/00031305.1991.10475783

Teerapabolarn, K. (2011). On the Poisson approximation to the negative hypergeometric distribution. *Bulletin of the Malaysian Mathematical Sciences Society. Second Series*, *34*(2), 331–336.

Thompson, S. K. (n.d.). Wiley: Sampling, 3rd Edition - Steven K. Thompson. Retrieved March 22, 2017, from http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470402318.html

Tohma, Y., Yamano, H., Ohba, M., & Jacoby, R. (1991). The Estimation of Parameters of the Hypergeometric Distribution and Its Application to the Software Reliability Growth Model. *IEEE Trans. Softw. Eng.*, *17*(5), 483–489. https://doi.org/10.1109/32.90450

Traat, I., & Ilves, M. (2007). The Hypergeometric Sampling Design, Theory and Practice. *Acta Applicandae Mathematicae*, *97*(1–3), 311–321. https://doi.org/10.1007/s10440-007-9119-9

Tripathi, R. C., & Gurland, J. (1977). A General Family of Discrete Distributions with Hypergeometric Probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(3), 349–356.

Walton, G. S. (1986). The Number of Observed Classes From a Multiple Hypergeometric Distribution. *Journal of the American Statistical Association*, *81*(393), 169–171. https://doi.org/10.2307/2287984

Wang, W. (2015). Exact Optimal Confidence Intervals for Hypergeometric Parameters. *Journal of the American Statistical Association*, *110*(512), 1491–1499. https://doi.org/10.1080/01621459.2014.966191

Whittaker, E. T., & Watson, G. N. (2010). *A Course of Modern Analysis*. Cambridge University Press.

Wright, T. (2012). *Exact Confidence Bounds when Sampling from Small Finite Universes: An Easy Reference Based on the Hypergeometric Distribution*. Springer Science & Business Media.

Wroughton, J. (n.d.). Distinguishing Between Binomial, Hypergeometric, and Negative Binomial Distributions | CAUSEweb. Retrieved February 1, 2017, from https://www.causeweb.org/cause/webinar/jse/2013-10

Wu, T. (1993). An Accurate Computation of the Hypergeometric Distribution Function. *ACM Trans. Math. Softw.*, *19*(1), 33–43. https://doi.org/10.1145/151271.151274

Zhang, H. (2010). A Note About Maximum Likelihood Estimator in Hypergeometric Distribution. *REVISTA COMUNICACIONES EN ESTADÍSTICA*, ([object Attr]). Retrieved from https://ideas.repec.org/a/col/000350/006974.html

Vita

After completing her work at Hudson High School, Lufkin, Texas, in 2011, Brooke Busbee entered Stephen F. Austin State University at Nacogdoches, Texas. She received the degree of Bachelor of Science with a minor in Secondary Education from Stephen F. Austin State University in May 2015. In August 2015, she entered the Graduate School of Stephen F. Austin State University, and received the degree of Master of Science in August 2017. During her time in graduate school, she worked as a graduate teaching assistant teaching two remedial math classes each semester.

Permanent Address: 402 Donna Dr.
Lufkin, TX 75904

This thesis uses the APA Style Manual.

This thesis was typed by Brooke Busbee.